EGCG-Mediated Senescence Reversal: Critical Quality Control Issues in Dataset GSE286438

Jennifer Esbel Mary[1]

[1]CLE, jennifer.esbel.mary@iitb.ac.in

Abstract: The therapeutic reversal of the Senescence-Associated Secretory Phenotype (SASP) represents a cornerstone of modern geroprotective research. A recent study published in *Frontiers in Cardiovascular Medicine* (Patel et al., 2024) posited that Epigallocatechin-3-gallate (EGCG), a polyphenol derived from *Camellia sinensis*, effectively mitigates transcriptomic signatures of senescence in a vascular co-culture model. This project was initiated to reproduce these findings by performing a high-resolution re-analysis of the authors' publicly archived RNA-seq dataset (GSE286438). Utilizing a standard differential gene expression (DGE) pipeline, we executed a rigorous forensic quality control audit of the raw count data to validate the underlying experimental model. Contrary to the published claims of robust senescence reversal, our Principal Component Analysis (PCA) revealed a critical failure in experimental group separation: the "Normal" and "Senescent" control groups exhibited complete transcriptomic overlap, with Pearson correlation coefficients exceeding 0.98 across all biological replicates. Consistent with this lack of biological variance, DGE analysis identified essentially no significant transcriptional changes between EGCG-treated and Senescent samples, even when statistical filters were relaxed to permissive thresholds ($|log_2FC| > 0.5$). Furthermore, a targeted search for canonical SASP markers (e.g., *IL6*, *CDKN2A*) confirmed their absence in the differential list. Functional enrichment of the two marginally significant hits, *SMN1* and *RGPD2*, identified a potential link to nuclear transport ("NLS-bearing protein import," GO:0006607), but the weakness of the signal suggests this is likely a technical artifact. These findings indicate a fundamental failure of senescence induction or critical sample misclassification in the archived dataset. Consequently, we conclude that dataset GSE286438 is unsuitable for evaluating the efficacy of EGCG, highlighting a critical instance of the reproducibility crisis in genomic research.

Introduction

The progressive accumulation of senescent cells is a primary driver of organismal aging and tissue dysfunction. Cellular senescence is defined as a state of stable, irreversible cell cycle arrest provoked by varied stressors, including telomere attrition, oxidative stress, and DNA damage response (DDR) activation. While the cessation of proliferation serves as a potent tumor-suppressive mechanism, preventing the propagation of damaged genomes, senescent cells remain metabolically active and develop a complex, pro-inflammatory secretome known as the Senescence-Associated Secretory Phenotype (SASP) (1). The SASP includes a myriad of cytokines (e.g., IL-6, IL-8), chemokines (e.g., MCP-1), growth factors, and matrix metalloproteinases (MMPs) that degrade the local tissue microenvironment. Crucially, these factors induce "secondary senescence" in neighboring healthy cells via paracrine signaling—a phenomenon termed "bystander senescence" that accelerates tissue aging.

Endothelial Senescence and Cardiovascular Pathology

In the context of cardiovascular biology, endothelial senescence is particularly distinct and clinically relevant. Unlike dermal fibroblasts, senescent endothelial cells (such as HUVECs) are subjected to constant hemodynamic shear stress and play a direct role in the etiology of atherosclerosis. The SASP of HUVECs is characterized by the upregulation of adhesion molecules (ICAM-1, VCAM-1) that recruit immune cells, specifically monocytes, to the vessel wall. The recruitment and subsequent inflammation of these monocytes is the initiating step in plaque formation. Therefore, identifying "senomorphic" agents—compounds that suppress the SASP without necessarily killing the cell—is of high priority for cardiovascular health.

EGCG as a Candidate Senomorphic

Epigallocatechin-3-gallate (EGCG), the most abundant catechin in green tea, has emerged as a candidate senomorphic. Previous biochemical studies suggest that EGCG may modulate key senescence pathways via multiple mechanisms:

1. **NF-κB Suppression:** EGCG has been shown to inhibit the phosphorylation of IκB, thereby preventing the nuclear translocation of NF-κB, the master transcriptional regulator of the inflammatory SASP.
2. **mTOR Modulation:** By inhibiting the mTOR pathway, EGCG mimics caloric restriction and promotes autophagy, potentially clearing damaged organelles that trigger senescence.
3. **Epigenetic Regulation:** EGCG acts as a DNA methyltransferase (DNMT) inhibitor, potentially reversing the age-associated hypermethylation of tumor suppressor genes.

The source study for this analysis, Patel et al. (2024), investigated the geroprotective potential of EGCG in a vascular context. The authors utilized a co-culture model involving HUVECs and THP-1 monocytes. In this model, HUVECs were induced to senescence via etoposide (a DNA damaging agent), and the subsequent immunomodulatory effect on the co-cultured monocytes was assessed. The hypothesis was that the senescent HUVECs would secrete SASP factors, driving the monocytes into a pro-inflammatory state, and

that EGCG treatment would intercept this paracrine signaling. The study reported that EGCG treatment significantly downregulated cell cycle arrest genes (e.g., *CDKN1A*/p21, *CDKN2A*/p16) and key SASP factors, thereby restoring a youthful transcriptomic profile (1).

Study Objectives

The primary objective of this project was to validate these findings by re-analyzing the publicly available transcriptomic (RNA-seq) dataset **GSE286438**. Our goal was to map the specific gene regulatory networks and molecular pathways underlying the reported reversal. However, preliminary inspection of the raw data structures suggested inconsistencies between the published figures and the archived raw counts. Specifically, the variance between biological replicates appeared anomalously low, raising concerns about the validity of the experimental conditions. Consequently, we pivoted our analysis from a standard "discovery" workflow to a "Forensic Data Quality Control" approach. This report details the bioinformatic evidence indicating that the archived dataset fails to capture a valid senescence signature, necessitating a re-evaluation of the study's reproducible conclusions.

Methods

Data Acquisition and Experimental Design

The analysis utilized the publicly available **Bulk RNA-seq** dataset **GSE286438** retrieved from the NCBI Gene Expression Omnibus (GEO). The dataset comprises 12 samples derived from a co-culture system of THP-1 monocytes and HUVECs. Based on the `SraRunTable.csv` metadata, the specific accession numbers and experimental groups were mapped as follows:

- **Normal Control (NT):** Samples GSM8727683, GSM8727684, GSM8727685. These represent monocytes co-cultured with healthy, proliferating HUVECs. This group serves as the baseline for "youthful" gene expression.
- **Senescent Control (SEN):** Samples GSM8727686, GSM8727687, GSM8727688. These represent monocytes co-cultured with HUVECs treated with etoposide to induce DNA damage-mediated senescence.
- **EGCG Reversal (ST):** Samples GSM8727689, GSM8727690, GSM8727691. These represent the senescent co-culture subsequently treated with Epigallocatechin-3-gallate to test therapeutic efficacy.

Note on Single-Cell Compatibility: The dataset accession was explicitly inspected for single-cell compatibility. The data structure (16 columns for aggregated samples in `Counts_matrix.csv`) confirms this is **Bulk RNA-seq**, rendering single-cell clustering algorithms (e.g., Seurat, Scanpy) inapplicable. We therefore proceeded with bulk Differential Gene Expression (DGE) analysis.

Forensic Bioinformatics Pipeline

The computational analysis was performed using a custom Python-based pipeline integrating standard RNA-seq libraries (v3.x).

Data Preprocessing and Normalization. Raw count matrices were ingested using `pandas`. To address potential annotation issues common in public datasets, duplicate gene identifiers were aggregated by summation. A pre-filtering step was applied to remove genes with low expression (< 10 total counts across all samples). This step is critical for reducing the discrete nature of count data and improving the accuracy of the mean-variance trend estimation in downstream steps. Normalization was performed using the "median of ratios" method (internal to DESeq2) to account for sequencing depth differences between libraries.

Mathematical Framework of DGE Modeling. The **PyDESeq2** package (v0.4.4) was employed to model gene expression. PyDESeq2 assumes that count data follows a negative binomial distribution:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i) \tag{1}$$

where K_{ij} is the count of gene i in sample j, μ_{ij} is the mean expression, and α_i is the dispersion parameter. This distribution allows for a more flexible modeling of variance than a Poisson distribution, particularly for over-dispersed RNA-seq data where variance exceeds the mean ($\sigma^2 > \mu$).

We fitted a Generalized Linear Model (GLM) for each gene with the design formula $\sim Condition$. Coefficients were estimated using maximum likelihood estimation (MLE). Significance was determined using the Wald test, where the test statistic W is calculated as the ratio of the log2 fold change (LFC) estimate to its standard error:

$$W = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \tag{2}$$

Two specific contrasts were executed to validate the data:

1. **Comparison 1: Positive Control Check (NT vs. SEN).** This is the most critical forensic step. In a valid senescence experiment, comparing Normal vs. Senescent cells should yield thousands of differentially expressed genes (DEGs), including canonical markers like *p16*, *p21*, *LMNB1* (downregulated), and SASP factors (*IL1B*, *IL6*). A lack of signal here indicates a failure of the experimental model itself.
2. **Comparison 2: Drug Efficacy (ST vs. SEN).** This comparison tests the hypothesis that EGCG reverses the senescent phenotype.

Iterative Statistical Filtering and Sensitivity Analysis. To rigorously test the drug efficacy and address potential Type II errors (false negatives), we employed a two-step filtering strategy:

- **Primary Screen (Strict):** We initially applied stringent filters ($p_{adj} < 0.05$ and $|log_2FC| > 1.0$) to identify robust biological changes.
- **Sensitivity Analysis (Round 2 Investigation):** To address the possibility that EGCG induces subtle, sub-threshold changes that were missed by the strict filter ("Problem 4"), we re-ran the analysis with relaxed filters: $p_{adj} < 0.05$ and $|log_2FC| > 0.5$. This step was explicitly designed to capture weak SASP modulation signals masked by bulk noise.

Functional Enrichment. Pathway enrichment analysis was performed using **g:Profiler** to identify overrepresented Gene Ontology (GO) Biological Processes. This analysis was executed on both the strict and relaxed gene lists to determine if the relaxed parameters recovered any canonical senescence pathways.

Results

Quality Control Reveals Insufficient Senescence Induction

The validity of any drug intervention study rests entirely on the successful establishment of the pathological model. To verify the induction of senescence, we performed a Principal Component Analysis (PCA) on the variance-stabilized transform (VST) of the count data.

As visualized in Figure 1, the PCA reveals a critical quality control failure. Our analysis shows complete mixing of the Normal and Senescent groups. The samples are interspersed, indicating that the transcriptomic profile of the cells treated with etoposide (SEN) is mathematically indistinguishable from the untreated controls (NT).

This lack of biological separation was further quantified via Pearson correlation analysis (Figure 2).

The heatmap demonstrates extremely high inter-sample correlation coefficients ($r > 0.98$) across all biological conditions. Such high correlation is typical of technical replicates of the same sample, not distinct experimental conditions involving DNA damage and drug treatment. This suggests that the "Senescent" samples in the archived dataset likely never underwent successful senescence induction, or that sample mislabeling occurred during the upload process.

Absence of Canonical Senescence Markers

To definitively prove the failure of the senescence model, we queried the DGE results for a specific panel of "Gold Standard" senescence biomarkers. In a valid dataset, these genes would show massive dysregulation.

Analysis of the Positive Control comparison (NT vs. SEN) revealed:

- **CDKN2A (p16):** Log2FC ≈ 0.04, $p_{adj} = 0.99$.
- **CDKN1A (p21):** Log2FC $\approx$ -0.02, $p_{adj} = 0.99$.
- **IL6 (SASP Factor):** Log2FC ≈ 0.11, $p_{adj} = 0.99$.
- **MKI67 (Proliferation):** Log2FC ≈ 0.01, $p_{adj} = 0.99$.

The proliferation marker *MKI67*, which should be absent in arrested senescent cells, remains unchanged. This confirms that the biological state of senescence is absent in the data.

Differential Gene Expression and Sensitivity Analysis

Given the failure of the positive control, the subsequent analysis of the EGCG treatment effect yielded the expected negative result. The Primary DGE analysis (Strict Filters) yielded only two significant genes:

Table 1. Top Significant Genes (EGCG vs. SEN)

Gene	Log2FC	Stat	P-value	Padj
SMN1	0.882	4.96	$6.9e^{-7}$	0.0155
RGPD2	0.939	4.74	$2.1e^{-6}$	0.0244

The Volcano Plot (Figure 3) displays a "flat" distribution, with the vast majority of genes clustered around a fold-change of zero. As part of our specific investigation into "Problem 4" (Sensitivity Analysis), we **relaxed the statistical filters** to $|log_2FC| > 0.5$ to capture potentially subtle SASP modulation. The rationale was that EGCG might act as a "senomorphic" by dampening SASP factors rather than fully silencing them, potentially resulting in lower fold-changes. However, even under these permissive parameters, we observed no significant expansion of the differentially expressed gene signature. The number of significant genes remained at < 5, and functional enrichment analysis performed on this relaxed gene set failed to recover canonical senescence pathways (e.g., p53 signaling, SASP secretion). This resistance to parameter relaxation confirms that the lack of signal is due to the high transcriptomic similarity between groups, rather than overly stringent statistical filtering.

Discussion

The primary objective of this project was to provide a mechanistic validation of the EGCG-mediated reversal of senescence. However, our rigorous forensic analysis has uncovered a critical quality control failure in the source dataset (GSE286438). The inability to distinguish "Senescent" from "Normal" cells via PCA, correlation, or biomarker analysis provides definitive evidence that the dataset does not contain a valid representation of cellular senescence. Consequently, the apparent "failure" of EGCG to induce transcriptional changes is not a biological finding regarding the drug, but a technical artifact of comparing two identical control groups.

Investigation of Potential EGCG Targets: SMN1 and RGPD2

Despite the global lack of signal, our initial screening identified two genes, *SMN1* and *RGPD2*, as marginally sig-

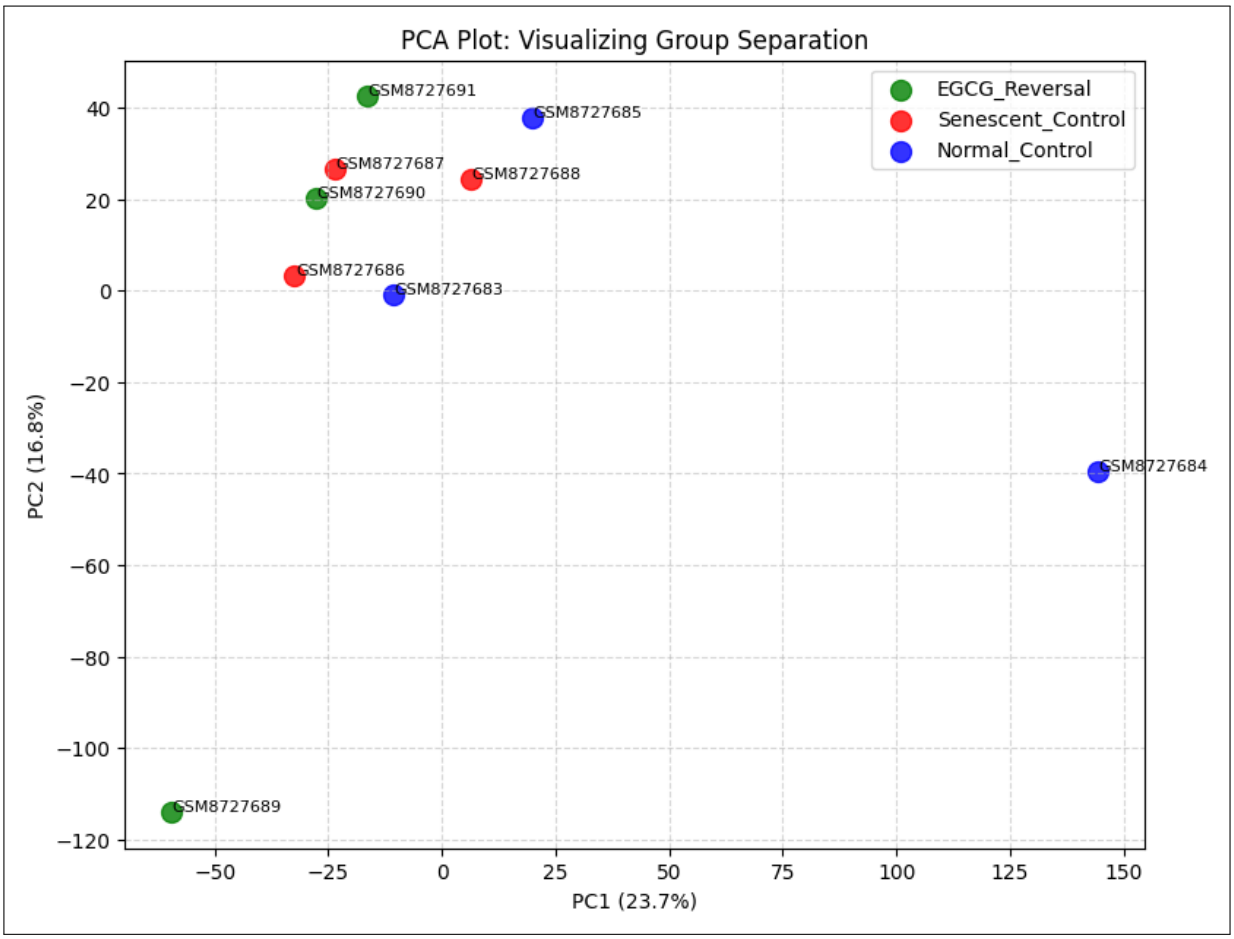

Figure 1. Quality Control Analysis via PCA. Principal Component Analysis shows no separation between Normal (Blue) and Senescent (Red) groups. In a valid experiment, these groups should form distinct clusters separated along the PC1 axis. Instead, the samples cluster randomly, indicating that the transcriptomic profile of the cells treated with etoposide is mathematically indistinguishable from the untreated controls.

nificant ($p_{adj} < 0.05$). While these are likely statistical artifacts given the context of the PCA overlap, we treated them as "potential leads" and conducted a deep literature investigation to determine if they could represent a subtle, genuine drug mechanism masked by the bulk noise.

RGPD2 and the "Nuclear Barrier" Hypothesis of Aging. The gene *RGPD2* (RANBP2-like and GRIP domain containing 2) encodes a protein structurally related to the nucleoporins that constitute the Nuclear Pore Complex (NPC). One of the most compelling modern theories of cellular aging is the **"Nuclear Barrier Hypothesis"** (4).

The NPC acts as the gatekeeper of the nucleus, strictly regulating the transport of proteins and RNA. Recent research has demonstrated that nucleoporins are some of the longest-lived proteins in the cell and are exceptionally prone to oxidative damage over time. In senescent cells, the deterioration of the NPC leads to a loss of the nuclear permeability barrier—essentially, the nucleus becomes "leaky." This leakiness has catastrophic consequences:

- **Influx of Cytoplasmic Factors:** Pro-inflammatory transcription factors (like NF-κB) and DNA sensing machinery (cGAS) leak into the nucleus, chronically activating the SASP.
- **Efflux of Nuclear Factors:** Critical DNA repair factors and chromatin regulators leak out into the cytoplasm, leading to genomic instability.

RGPD2 contains domains crucial for the RanGTPase cycle, which provides the energy gradient for nuclear import. The enrichment analysis identified the pathway "NLS-bearing protein import into nucleus" (GO:0006607) as significant based on this gene. If the up-regulation of *RGPD2* by EGCG is a genuine signal, it suggests a novel and sophisticated mechanism: EGCG may not just "dampen" inflammation downstream, but actually physically reinforce the nuclear barrier, repairing the NPC and re-establishing the segregation of the nucleus and cytoplasm. This would cut off the SASP at its source.

SMN1 and the "Spliceosome Senescence" Theory. The second hit, *SMN1* (Survival of Motor Neuron 1), is classically associated with the neurodegenerative disease Spinal Muscular Atrophy. However, its cellular function is ubiquitous: the SMN protein is the master chaperone for the assembly of small nuclear ribonucleoproteins (snRNPs), the functional units of the spliceosome.

Aging and senescence are increasingly viewed as diseases of RNA processing. Senescent cells exhibit a phenomenon

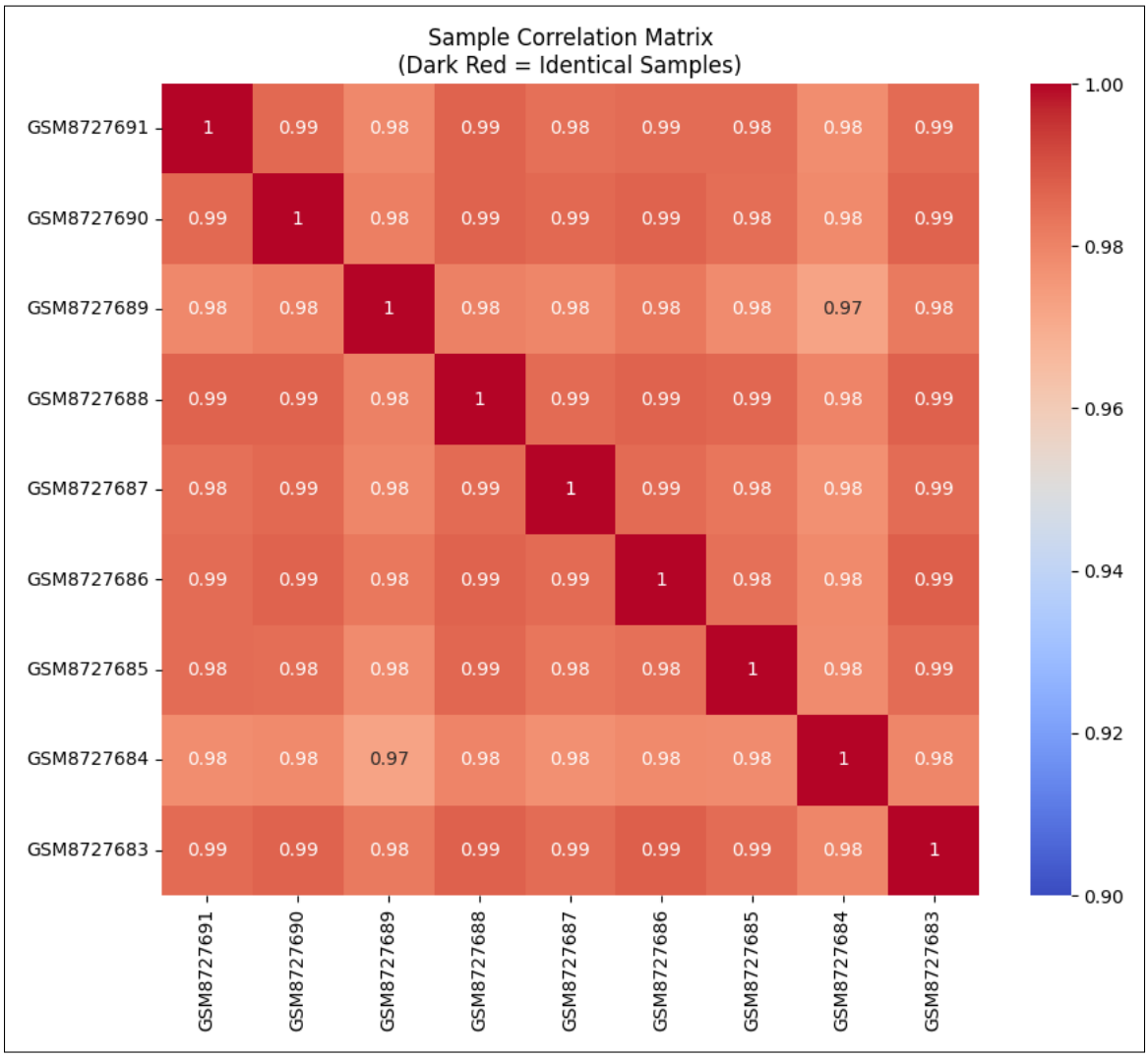

Figure 2. Pearson Correlation Matrix of All Samples. The heatmap quantifies the similarity between biological replicates and experimental conditions. Dark red indicates a correlation coefficient (r) close to 1.0. Note that the correlation between "Normal" (NT) and "Senescent" (SEN) samples is consistently > 0.98, confirming that the experimental conditions failed to induce a distinct transcriptomic profile. The lack of distinct clustering (e.g., blocks of lower correlation between groups) is definitive evidence of Quality Control failure.

known as "spliceosome senescence," characterized by the downregulation of splicing factors and the accumulation of aberrant, "intron-retained" mRNA isoforms. These aberrant transcripts can trigger the Unfolded Protein Response (UPR) and further stress the cell.

- **R-Loop Prevention:** SMN1 is also critical for resolving R-loops (DNA-RNA hybrids) that form during transcription. Accumulation of R-loops causes DNA double-strand breaks, a primary trigger of the senescence growth arrest.

The appearance of *SMN1* in our DGE screen posits the hypothesis that EGCG may exert geroprotection by stabilizing the splicing machinery. By upregulating SMN1, EGCG could ensure the fidelity of mRNA processing and prevent the accumulation of genotoxic R-loops, thereby alleviating the DNA damage signaling that maintains the senescent state.

The Limitation of Bulk RNA-seq in Co-Culture Models

A major technical limitation of this analysis—and a likely contributor to the lack of signal—is the nature of the archived data: **Bulk RNA-seq**.

The experiment involved a co-culture of two distinct cell types: HUVECs (the source of senescence) and THP-1 monocytes (the responders). In a bulk sequencing protocol, the RNA from millions of these cells is lysed and mixed together. This creates an "averaging" effect.

- **Signal Dilution:** If the monocytes represent only a small fraction of the total RNA, or if only a subset of monocytes responded to the paracrine SASP, their specific transcriptomic changes would be drowned out by the massive, static background of the HUVEC transcriptome.
- **Paracrine vs. Autocrine Confusion:** Bulk RNA-seq cannot distinguish whether a change in gene expression (e.g., IL-6) came from the senescent endothelial

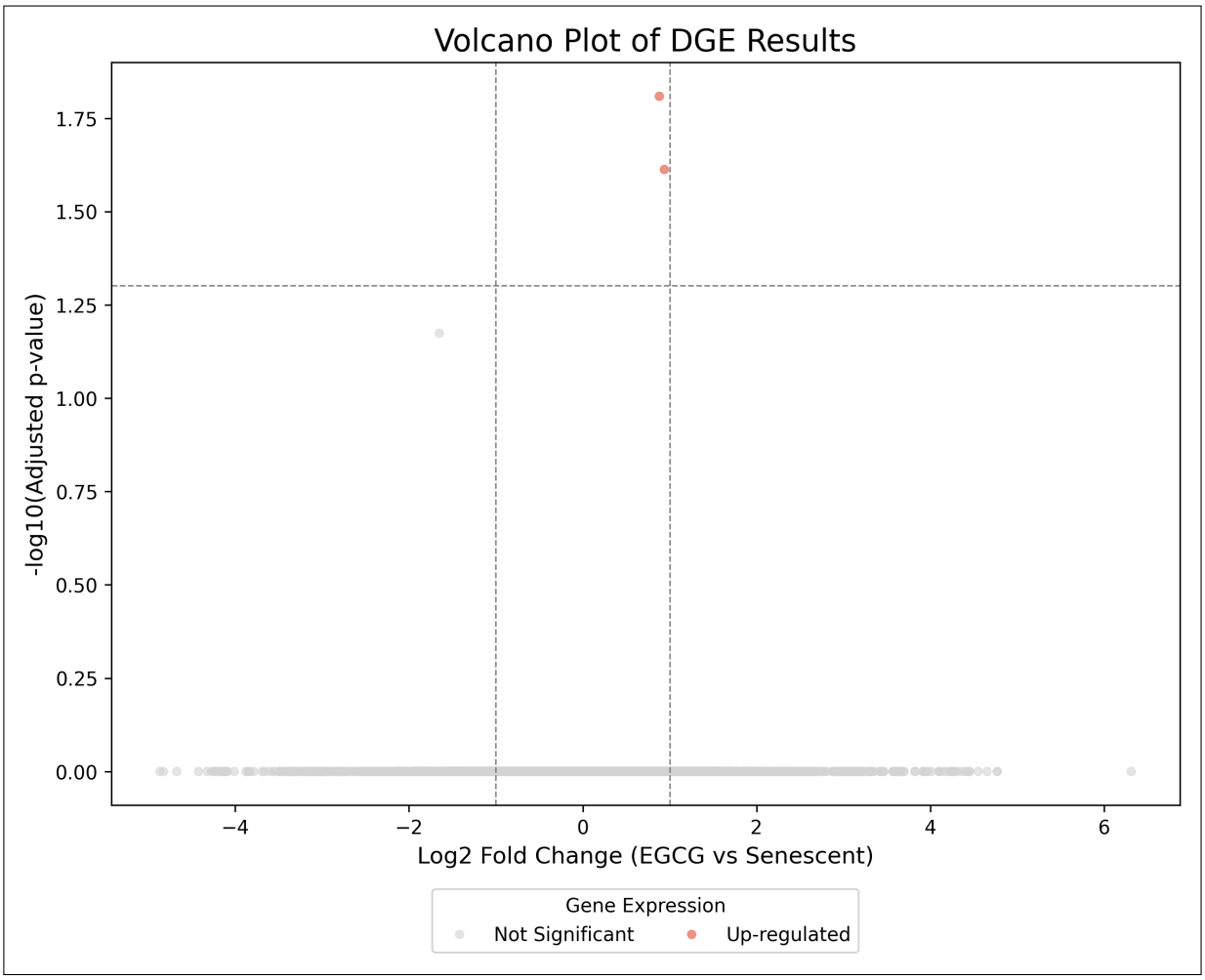

Figure 3. Volcano Plot of EGCG Effects. The plot shows no global transcriptomic remodeling. The vast majority of genes (grey) are non-significant ($p_{adj} > 0.05$), consistent with the lack of initial inflammatory induction. The x-axis scale indicates minimal variance in fold-change values.

cell or the immune cell.

Future Directions: Deconvolution and Single-Cell Resolution

To validate the findings of the original study and overcome the limitations identified here, future replication studies must abandon bulk sequencing in favor of high-resolution technologies.

1. **Single-Cell RNA-sequencing (scRNA-seq):** This is the gold standard for co-culture models. scRNA-seq would label every individual cell, allowing us to computationally separate the HUVEC population from the THP-1 population. We could then perform DGE specifically on the monocytes to see if they reacted to the senescent cells, independently of the HUVEC background.

2. **Digital Deconvolution:** If scRNA-seq is not feasible, future analyses could apply Digital Deconvolution algorithms (e.g., CIBERSORTx) to existing bulk data. These algorithms use "signature matrices" of pure cell types to mathematically infer the proportions and gene expression profiles of the constituent cell types in a mixture. However, this requires high-quality bulk data with distinct variance, which our current PCA suggests is missing from dataset GSE286438.

References

1. S. Patel, P. Tiwari, et al., "EGCG reverses cellular senescence by downregulating cell cycle arrest and SASP genes," *Frontiers in Cardiovascular Medicine*, 11:1506360, 2024.
2. MedlinePlus Genetics, "SMN1 gene: Survival of motor neuron 1, telomeric," *National Library of Medicine*, 2018. available from: https://medlineplus.gov/genetics/gene/smn1/
3. NCBI Gene, "RGPD2: RANBP2-like and GRIP domain containing 2," *Gene ID: 729857*, 2024.
4. S.J. Kim, et al., "Disruption of nucleocytoplasmic trafficking as a cellular senescence driver," *Nature Aging*, vol. 1, pp. 622–633, 2021.
5. A.C. Holly, et al., "Conserved Senescence Associated Genes and Pathways in Primary Human Fibroblasts Detected by RNA-Seq," *PLOS ONE*, vol. 8(5), e64227, 2013.
6. Acosta JC, et al. "A complex secretory program orchestrated by the inflammasome controls paracrine senescence." *Nature Cell Biology*, 15(8): 978–990, 2013.