Gene Regulatory Network Inferences using transformer-based methods and scGPT

Name: Chone Makhubela

Bachelor of Medical Science (Hons) in Bioinformatics

Institution: University of Cape Town

Supervisor: Dr Musalula Sinkala

Co-Supervisor: Emmanuel Biryabarema

Year: November 2025

Declaration

I, Chone Makhubela, declare that work entitled "*Gene Regulatory Network Inferences using transformer-based methods and scGPT"* is my own work, conducted in fulfilment of the requirements for the degree of Bachelor of Science (Honours) in Bioinformatics at the University of Cape Town. This work has not been submitted before for any other degree or examination at this or any other institution. All sources used have been acknowledged and referenced in accordance with the required academic conventions. Artificial Intelligence (AI) was implemented to modify the report into an academic tone only.

Chone Makhubela

Signature

Abstract

This project addresses the challenge of accurately inferring Gene Regulatory Networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data by comparing the performance of two transformer-based foundation models against one another. Traditional GRN inference struggles with the sparsity and high dimensionality of single-cell data, limiting the reliable determination of regulatory directionality. We leveraged the Single-cell Generative Pre-Trained Transformer (scGPT) foundation model, which treats the transcriptome as a language, to overcome these limitations. The project utilized the Immune All Human scRNA-seq dataset and employed a dual-pipeline approach, comparing the scGPT-based gene program identification with the Single-cell Graph-based Regulatory Elements Analysis Toolkit (scGREAT) transformer model for Consensus GRN inference.

The scGPT analysis successfully identified six distinct, lineage-specific gene programs. For example, Gene Program 1 was highly activated in CD14+ Monocytes but absent in Plasmacytoid Dendritic Cells, while Gene Program 2 was prominent in Plasma Cells and CD20+ B Cells. Functional enrichment analysis provided a unified mechanistic explanation, strongly converging on the regulation of cellular fate through the Ubiquitin-Proteasome System (UPS) and the induction of Cellular Senescence as key regulatory axes. The most significant enriched biological process was the positive regulation of cellular component movement ({-log} _ {10}P \approximately 3.0).

On the contrary, the scGREAT Consensus GRN identified highly interconnected regulatory hubs that likely represent the master coordinators such as Deubiquitinates (DUBs) responsible for integrating DUB activity with fundamental immune processes like cell trafficking and longevity.

These findings demonstrate that transformer-based models can move beyond descriptive characterization to offer mechanistic hypotheses, positioning the identified regulatory hubs and UPS components as novel, high-priority therapeutic targets for modulating immune function in the context of chronic disease, aging, and immunotherapies. The work lays the foundation for targeted experimental validation to translate these computational predictions into functional biological insights.

Table of Contents

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Musalula Sinkala and co-Supervisor, Emmanuel Biryabarema, for their guidance, support, and valuable feedback throughout this research. Their expertise was essential in the completion of this study on Gene Regulatory Network inference using transformer-based models and scGPT.

I also thank Inter-University Institute for Data Intensive Astronomy (ilifu) for providing access to their cluster and jupyter environment that made this work possible.

Introduction

Gene Regulatory Networks (GRNs) are complex biological systems that facilitate the understanding of cellular function and mechanisms, gene to gene interactions, disease progression, and the identification of potential therapeutic targets. These networks illustrate interactions between transcription factors, co-factors, and their target genes enabling us to understand how they mediate fundamental biological processes (Cui et al., 2024; Wang et al., 2024).

The advent of single-cell RNA sequencing (scRNA-seq) has revolutionised transcriptomic analysis by enabling gene expression profiling at unprecedented cellular resolutions. However, traditional methods of inferring GRNs struggle with high dimensionality, data sparsity, and the inability to reliably determine the directionality of regulatory interactions (Afonja et al., 2024). These older models were primarily designed for specific tasks and lacked the holistic, contextual understanding needed to accurately model complex regulatory cascades across the entire cellular system.

In response, the field has seen a shift toward foundation models, inspired by the success of the transformer architecture in Natural Language Processing (NLP). This approach treats the transcriptome as a language, allowing models like scGPT to leverage self-supervised pre-training on millions of cells to develop a deep, contextual understanding of gene-gene relationships (Cui et al., 2024).

This project aims to bridge a significant research gap by conducting a systematic, head-to-head comparison of Gene Regulatory Networks inferred by the foundation model scGPT against other transformer-based methods, such as scGREAT. Specifically, this research will evaluate the accuracy of these inferences and determine how these cutting-edge models can assist in identifying potential drug targets, providing an evidence-based roadmap for their effective application in personalized medicine (Afonja et al., 2024; Wang et al., 2024a).

Literature Review

Gene Regulatory Networks (GRNs) are biological networks that aid in the understanding of the function and the mechanism of cells, the interactions between genes, disease progression, and identifying drug targets (Aibar et al., 2017). The transcription factors, co-factors, and target genes in these networks mediate fundamental biological processes and enable us to understand interactions between genes. The sparsity of single cell RNA sequencing data, and the limited sequencing capacity and depth, makes the use of foundational transformer models hard (Cui et al., 2024). The traditional models currently used in single-cell research are only designed for specific tasks and are seldom used in other downstream tasks such as GRNs, which is what this research aims to explore.

Various traditional methods of inferring gene regulatory networks solely rely on static gene expression and pseudo-time estimation to create causal graphs (Akers & Murali, 2021). The use of scGPT as a foundational model in biology is underexplored due to the large amounts of single-cell data required to train the model. However, scGPT's optimized generative pre-training uses the gene symbols to encode the relationships in its gene embeddings. This enables the model to accurately learn the data and make predictions or capture insightful biological meaning and function when a new dataset is introduced (Cui et al., 2024). This research aims to understand if scGPT can make accurate inferences on gene regulatory networks when compared to other transformer-based methods, and how these inferences can assist in identifying potential drug targets.

The Evolution of Gene Regulatory Network Inferences

Traditional Methods

Early on, scientists used correlation-based methods like GENIE3, which were effective at finding genes that behaved similarly (Huynh-Thu et al., 2010). However, these tools could not distinguish between a direct causal relationship and a shared regulatory influence from a third gene, often leading to false positives. To address this, more precise information theory-based methods, such as Protocol Independent Detection and Classification (PIDC), were developed. These tools were better at finding complex, non-linear relationships and were less susceptible to simple co-expression patterns (Chan et al., 2017). Still, a key limitation was their computational inefficiency and an inability to reliably determine the directionality or relative influence of genes within a regulatory interaction.

Later, more advanced machine learning methods were introduced. Scene-aware Semantic Navigation with Instruction-guided Control (SCENIC), for example, was a two-step process that leveraged extra biological data specifically, transcription factor binding motifs to validate and refine the inferred connections, making the networks more biologically plausible (Aibar et al., 2017). Other early AI models, like variational autoencoders, aided in making sense of the massive amount of single-cell data by learning a latent, compressed representation of gene expression (Wang et al., 2024). Despite these advancements, these methods still had a major flaw of often missing the big picture, failing to capture how genes work together across the entire cellular system. This is the gap that new foundational, transformer-based models are trying to fill by learning a more holistic and context-aware representation of gene interactions.

Single-cell transcriptomics

The field of single-cell transcriptomics traces its origins to the 1990s with pioneering work that enabled the analysis of gene expression in individual cells (Aldridge & Teichmann, 2020). Traditional methods mainly used in this area, such as Quantitative Polymerase Chain Reaction (qPCR), could only profile a smaller number of genes, limiting large-scale discovery. A significant breakthrough came with the first publication of single-cell RNA sequencing (scRNA-seq), which involved profiling rare cell types during mouse development and was motivated by the sparsity of the cells (Tang et al., 2009). Single-cell transcriptomics holds immense potential to enhance the understanding of health and disease.

To advance the field, technology had to be scaled to fit larger amounts of cells and to be able to profile many of these cells in parallel (Aldridge & Teichmann, 2020). This led to the creation of models like Single-cell Tagged Reverse Transcription sequencing (STRT-seq) (Islam et al., 2011) and Cell Expression by Linear amplification and sequencing (CEL-seq) (Hashimshony et al., 2012). Full-length methods like Switching Mechanism at the 5′ end of RNA Template sequencing (SMART-seq) and its successor, SMART-seq2, used unique techniques for RNA amplification. High-throughput methods utilizing microfluidics, such as Drop-seq (Macosko et al., 2015), dramatically increased the scale to thousands of cells.

More recently, multi-modal single-cell methods were developed to measure and combine information from different molecules such as RNA, DNA, and proteins (Aldridge & Teichmann, 2020). These allow scientists to study complex regulatory and communication networks between cells in much greater detail. Alongside these laboratory advances, new

computational tools were developed for processing and analysing single-cell data, helping identify cell types and states, find marker genes, and track developmental paths. The newer experimental techniques, including those that integrate multiple layers of biological data and spatial methods, continue to drive the development of advanced computational approaches (Aldridge & Teichmann, 2020).

Methods of analysing gene expression are rapidly emerging, with an exciting use of single-cell transcriptomics being the improvement of lab-grown cell models like organoids. It is now clear that this technology can be applied in many important medical areas, such as cell-based models, cell therapies, regenerative medicine, and finding new drug targets (Aldridge & Teichmann, 2020).

The Dawn of Foundation Models

The limitations of methods that were used before foundational methods were introduced have paved the way for a new shift in single-cell analysis, one inspired by the immense success of foundation models in natural language processing (NLP) (Xie et al., 2024). The core idea is to draw a powerful analogy about a cell's transcriptome, or its complete set of gene expressions, which can be viewed as a sentence, with each individual genes acting as a word in that sentence (Cui et al., 2024). A transformer-based model which is a deep learning model that is designed to process data in a sequence and learn the relationships between those sequences without considering the distance between them. Just as a language model learns grammar and context by processing vast amounts of text, transformer models can learn the rules of cellular biology by training on millions of cell sentences from a large pool of unlabelled single-cell data. This self-supervised approach allows the model to develop a deep, contextual understanding of gene-gene relationships without needing explicit labels (Cui et al., 2024; Theodoris et al., 2023).

The Rise of Transformer Models for single-cell Analysis

scGPT: a Generative Foundation Model

scGPT represents a landmark in this new paradigm. It is a generative pre-trained transformer model that has been trained on a massive amount of single-cell data (Cui et al., 2023). Its primary objective during pre-training is to predict masked gene expression values, a task that forces the model to learn the intricate dependencies between genes (Cui et al., 2024). By training on a diverse and extensive dataset, scGPT learns the embeddings as a representation

of genes and cells, which can be thought of as a rich, numerical summary of their biological state.

Crucially, scGPT's attention mechanism and learned embeddings are central to its ability to infer regulatory interactions. Within the transformer architecture, the attention mechanism calculates attention scores which are numerical values that measure how strongly one input element should focus on another element when information is being processed and counts how much each gene attends to every other gene in each cell. High attention scores between a transcription factor and a target gene can be interpreted as a strong potential regulatory link, providing a new way to infer GRNs that captures subtle, context-specific interactions that traditional methods might miss. These learned embeddings and attention patterns can then be extracted and analysed to build a novel, data-driven gene regulatory network, which is the key focus of this project.(Cui et al., 2024)

scGREAT: A Supervised Deep Learning Model

The development of scGREAT, a transformer-based model for GRN inference, is an important step forward in single-cell transcriptomics research. Unlike older statistical or machine learning methods, scGREAT uses deep learning and attention mechanisms to capture complex relationships between genes, including long-range interactions that are often missed by traditional tools(Wang et al., 2024). This makes the model well-suited for analysing large and noisy single-cell RNA-seq datasets, which continue to grow rapidly as new technologies improve.

In comparisons with other state-of-the-art GRN inference methods, scGREAT showed higher accuracy and better reliability across different datasets. It not only predicted gene interactions more effectively but also identified biologically meaningful regulatory links supported by existing studies(Wang et al., 2024). By combining advanced deep learning techniques with biological knowledge, scGREAT offers both strong performance and useful insights. This makes it a promising tool for studying gene regulation and for applications such as disease modelling, drug target discovery, and understanding complex cellular systems(Wang et al., 2024).

Traditional Vs. Foundation Models

When it comes to accuracy and biological interpretability, foundation models like scGPT demonstrate a clear advantage. By learning from a vast, unlabelled dataset, these models develop a rich, contextual understanding of the entire transcriptome, allowing them to capture complex, non-linear relationships that often elude simpler, correlation-based methods (Cui et al., 2024). Studies have shown that when validated against ground-truth data from perturbation experiments, models like scGPT can achieve state-of-the-art performance on a variety of downstream tasks, including Gene Regulatory Network inference (Cui et al., 2024; Theodoris et al., 2023).

A key to their success lies in the attention mechanism, which provides a novel avenue for interpretability. The attention scores can be extracted and analysed to understand which genes the model considers most important for predicting a target gene's activity. This appears to be a more nuanced and biologically meaningful insight into regulatory dynamics than the simple co-expression links offered by traditional methods (Cui et al., 2024).

However, the powerful capabilities of these models come with notable limitations, particularly in scalability and computational cost. Traditional methods such as GRNBoost2 are praised for their computational efficiency and ability to scale to large datasets (Moerman et al., 2019). In contrast, the self-attention mechanism in transformer models has a computational complexity that increases quadratically with the number of genes, making it computationally expensive and memory-intensive to process very large datasets (Cui et al., 2024). This high computational demand and the need for powerful Graphics Processing Unit (GPU) resources are significant practical barriers for many research labs.

Furthermore, a critical evaluation of foundation models must consider their inherent limitations and potential biases. When the training data under-represents certain cell types or disease states, the model may learn spurious correlations and exhibit biases, leading to inaccurate predictions in those specific contexts (Cui et al., 2024). Additionally, while the attention mechanism offers some interpretability, the overall complexity of a large transformer model can still make it difficult to fully understand the underlying biological mechanisms driving its predictions.

In summary, while traditional methods offer a degree of simplicity and scalability, foundation models like scGPT provide superior performance and biological interpretability by learning a

deeper, contextual understanding of the cell's regulatory landscape. This project will evaluate to what extent these tools can be effective in healthcare, especially in identifying drug targets.

Translating Networks into Therapeutic Insights

GRNs In Drug Target Identification

Understanding how genes interact is a crucial step in discovering new drug targets. An accurate GRN acts like a detailed road map of a cell, showing which genes are more influential than the others. By using this map, we can predict what will happen if we disrupt a specific gene by targeting it with a drug. The goal is to find a gene that, when switched off, can stop or slow down a disease process without causing too many side effects. This is the ultimate objective of drug target identification, and accurate GRN inference is the key to unlocking it (Aalto et al., 2020).

The scGPT Advantage

Foundation models like scGPT offer a significant advantage over traditional methods in this area due to their deep understanding of the cellular context. Traditional models often build a network based on simple correlations or a limited number of known interactions. This can make them less effective at predicting the broad, cascading effects of a drug on a cell's entire network (Rossner et al., 2025).

In contrast, scGPT's pre-training on millions of cells allows it to learn the complex web of gene-gene relationships. This means that when we want to predict the effect of a new drug, scGPT can provide a more holistic picture. It can use its learned representations to simulate how a drug targeting one gene might ripple through the entire network, affecting other genes and pathways that were not directly targeted. This capability is especially powerful for complex diseases like cancer, where a single drug can have many downstream effects (Cui et al., 2024). By using scGPT, we can not only identify a potential drug target but also better predict the full range of its therapeutic and side effects before ever needing to run an expensive and time-consuming experiment.

In essence, scGPT and similar models are not just building better maps but aid in testing the effects of a drug on an entire cellular network, a task that is practically impossible with traditional methods (Cui et al., 2024; Rossner et al., 2025).

The scGREAT Advantage

The computational tool scGREAT significantly advances single-cell research by specializing in accurate Gene Regulatory Network (GRN) inference from complex single-cell transcriptomics data, a process essential for understanding cellular behaviour and molecular mechanisms (Wang et al., 2023). scGREAT's primary strength lies in its novel transformer-based deep-language model architecture, which is trained by integrating gene expression data from single cells with external gene biotext information (Wang et al., 2024). This unique architectural approach allows scGREAT to effectively handle the high sparsity and noise typical of single-cell RNA-seq data, outperforming previous state-of-the-art methods in accurately identifying regulatory relationships between transcription factors and their target genes (Wang et al., 2024). By inferring these high-confidence GRNs, scGREAT enables researchers to gain mechanistic insights into cell fate determination, offering superior resolution for downstream analyses such as heterogeneity analysis and the inference of regulation-based pseudo-time trajectories (Badia-i-Mompel et al., 2023; Wang et al., 2023).

Conclusion & Future directions

In summary, the journey of GRN inference has evolved from simple statistical associations to complex, data-driven foundation models. Traditional methods, like correlation-based tools and early machine learning algorithms, provided foundational insights but were ultimately limited by their inability to capture the full, contextual complexity of the transcriptome. They often struggled to distinguish direct from indirect relationships and lacked a holistic view of the cell's regulatory landscape. The arrival of transformer-based models, exemplified by scGPT, marks a new era. By treating the transcriptome as a language, these models leverage self-supervised learning on massive datasets to develop a deep, contextual understanding of gene-gene relationships, offering superior accuracy and a more nuanced form of interpretability through their attention mechanisms.

The Research Gap and Project Rationale

Despite the promise of these new AI models, a significant research gap remains. While scGPT has demonstrated impressive performance on various tasks, a systematic, head-to-head comparison of Gene Regulatory Networks inferred by scGPT against those from established traditional methods, specifically for the task of identifying drug targets, is currently underexplored. Previous studies have shown that the performance of GRN inference

methods is highly variable, and it remains unclear how the unique capabilities of a foundation model translate to a direct, real-world application like drug target discovery. This project is necessary to bridge this gap. By critically evaluating these methods within the context of identifying drug targets, this research will provide a clear and evidence-based roadmap for when and where these powerful new tools are most effective. Ultimately, the findings will contribute to the field by providing a critical evaluation of a cutting-edge AI-driven method, guiding future research and development in drug target discovery and personalized medicine.

Materials and Methods

This project involved secondary data analysis based on studies by Cui et al. (2024) and Wang et al. (2024). All analyses were conducted on the Ilifu high-performance computing platform using python programming Language v3.13.2. The workflow adopted for our analysis is outlined in the steps below.

Data Acquisition and Preprocessing

The Immune All Human scRNA-seq dataset was sourced from the GitHub repository (Cui et al., 2024 that contains the code used in inferring GRNs and subsequently pre-processed using the scGPT model's internal pipeline, which consisted of three sequential steps. The initial step was filtering, which removed low-quality data by setting a minimum gene count threshold. Next, normalization was performed, where gene expression values were adjusted based on the total read count per cell to account for library size differences. The final step involved feature selection, where Highly Variable Genes (HVGs) were selected to retain the most biologically informative genes for subsequent analysis.

scGPT-Based Gene Program Inference

Model Initialization and Feature Extraction

The pretrained scGPT foundation model (Cui et al., 2024) was loaded into a Jupyter Notebook environment. This model incorporates a Gene Encoder for discrete gene identities, a Value Encoder for continuous expression levels, and a 12-layer Transformer Core responsible for learning complex gene-gene dependencies. Data-independent gene embeddings were extracted from the model's representations. A filtering step was applied to select genes present in both the dataset's HVGs and the scGPT model's comprehensive vocabulary (>30,000 genes), resulting in a final set of 1,200 genes for embedding analysis.

Clustering and Network Construction

The retrieved gene embeddings were used to construct a gene embedding network. Louvain clustering was performed on this network to partition genes into functional groups, or gene programs. To ensure biological relevance and stability, only clusters containing five or more genes were retained for subsequent analysis. A Cosine Similarity Network was then calculated between the identified gene programs, where edges represented the functional proximity of the programs, allowing for visualization of inter-program relationships.

scGREAT-Based Consensus GRN Inference

The scGREAT pipeline, which adapts the Transformer architecture for GRN inference, was applied to the same Immune Human Dataset. The Biobert model was utilized to generate numerical embeddings for 12,303 unique gene names in batches of 128, resulting in a (12,303, 768)-dimensional embedding matrix.

Sparse Network Generation and Correlation

An initial, sparse Gene Regulatory Network (GRN) was constructed using the k-Nearest Neighbours method, with k=50 to limit density. This generated an undirected graph of 12,303 nodes and 474,834 edges based on cosine similarity. Subsequently, an Expression Correlation Matrix was calculated using the Spearman method for the top 1,000 most active genes.

Consensus Network Filtering

A Consensus GRN was generated by combining the sparse network with the expression correlation data. To refine the visualization and focus on high-confidence interactions, a rigorous filtering step was performed, retaining only the top 5% of all edges based on regulatory weight. This final network was used for identifying and visualizing the Top 200 central genes (hubs).

Pathway and Functional Enrichment Analysis

Pathway enrichment analysis was performed on the identified gene programs and regulatory hubs. Databases queried included Reactome, Gene Ontology (GO) Biological Process (GO: BP), and GO Molecular Function (GO:MF) to identify the known biological pathways and molecular roles associated with the gene set. Statistical significance was reported as the {-log} _ {10} ({Adjusted P-value}).

Results

scGPT Gene Program Identification and Activation

Following preprocessing, gene embeddings were successfully retrieved for 1,200 genes. Louvain clustering identified a set of six distinct gene programs (clusters 0-5), each

comprising five or more genes. A Hierarchical Clustering Heatmap (Figure 1)

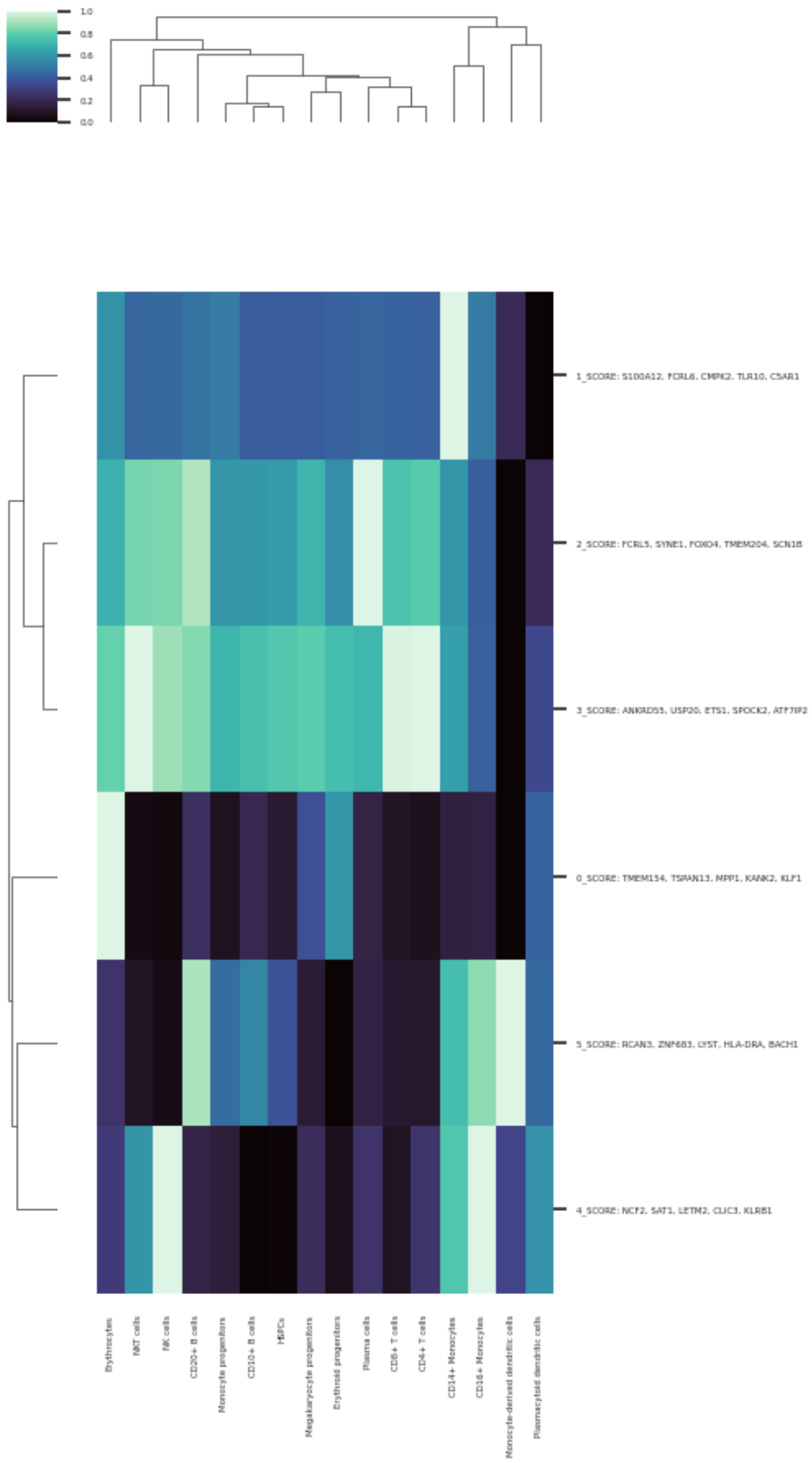

Figure 1. *Hierarchical Clustering of Gene Expression Signatures in Human Immune and Hematopoietic Cells. A heatmap displaying the relative expression scores of six clustered*

gene modules (rows) across fifteen sorted immune and hematopoietic cell populations (columns). Both rows and columns are hierarchically clustered (dendrograms) based on expression profile similarity. The colour bar indicates normalized expression scores from 0.0 (low, black/dark purple) to 1.0 (high, bright aqua/green). Distinct signatures are visible, including the high expression of Scores 1 and 2 across most cell types and the restricted expression of Score 4 to the Erythrocyte-Megakaryocyte progenitor lineage.

visualized the activation patterns of the identified gene programs across 16 human immune cell samples, revealing specific lineage-restricted expression. Gene Program 1 (comprising *S100A12*, *FCRL6*, *CMPK2*, *TLR10*, and *C5AR1*) displayed high activation in CD14+ Monocytes but was entirely absent in Plasmacytoid Dendritic Cells. Conversely, Gene Program 2 (*FCRL5*, *SYNE1*, *FOXO4*) exhibited high expression in Plasma Cells and CD20+ B Cells yet showed no detection in Monocyte-derived Dendritic Cells. Lastly, Gene Program 0 (*TMEM154*, *TSPAN13*, *MPP1*) was highly activated exclusively in Erythrocytes, strongly suggesting a specific, non-immune lineage function.

The relationships between these programs were visualized in a Cosine Similarity Network (Figure 2)

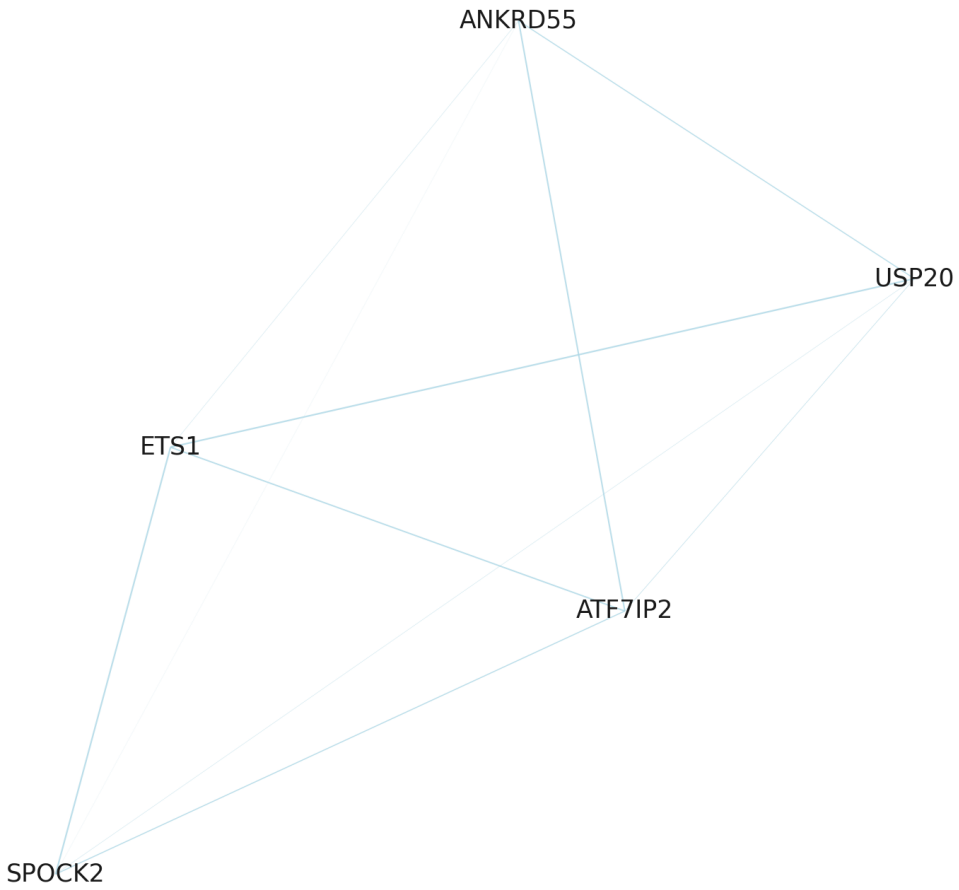

Figure 2. *Network Analysis Highlighting Interactions Among Key Genes. The figure illustrates the connectivity among five genes (ANKRD55, USP20, ETS1, SPOCK2, and ATF7IP2). It is displayed as a fully connected graph where the thickness or intensity of the edges (lines) reflects the strength of the relationship (Pearson correlation coefficient), showing that while all genes are interconnected, some pairs have a stronger co-expression link than others.*

which showed the interconnectedness of the identified gene clusters, confirming functional proximity between programs such as {ANKRD55, USP20, ETS1, SPOCK2, ATF7IP2}.

Functional Enrichment Analysis

Pathway enrichment analysis provided biological context for the identified gene programs, with consistent results across multiple databases.

Reactome Pathway Analysis

The Reactome 2022 analysis (Figure 3) highlighted pathways primarily associated with cellular aging and protein regulation. The top enriched pathways were Oncogene Induced Senescence ({-log} _ {10}P \approximately 1), Ub-specific Processing Proteases, and Cellular Senescence. These findings suggest that the gene programs are strongly involved in mechanisms regulating protein stability and cellular fate decisions.

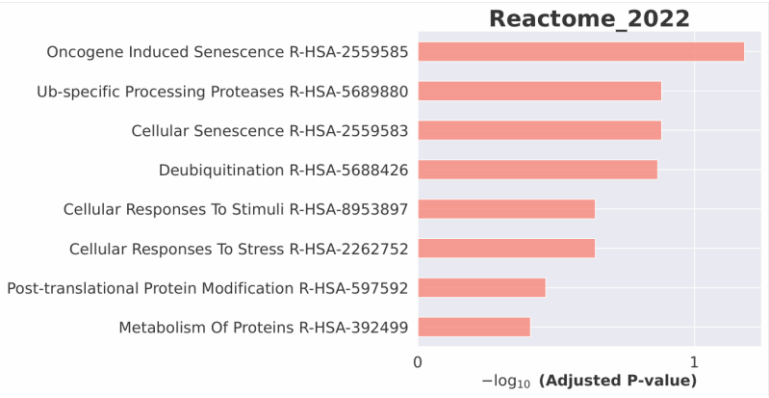

Figure 3. *Reactome pathway enrichment analysis of differentially expressed genes. The bar plot illustrates the top enriched pathways identified from the Reactome 2022 database based on adjusted p-values (displayed as –log₁₀ values). The most significantly enriched pathways include "Oncogene Induced Senescence" (R-HSA-2559585), "Ub-specific Processing Proteases" (R-HSA-5689880), and "Cellular Senescence" (R-HSA-2559583).*

Gene Ontology (GO) Molecular Function Analysis

The GO Molecular Function 2021 analysis (Figure 4) indicated a predominant role in enzymatic regulation. The highest significance was observed for terms related to protease activity and inhibition (*metalloendopeptidase inhibitor activity* and *cysteine-type peptidase activity*). Furthermore, highly significant terms such as *thiol-dependent deubiquitinase* and *deubiquitinase activity* confirmed the gene set's strong involvement in the ubiquitin-proteasome system.

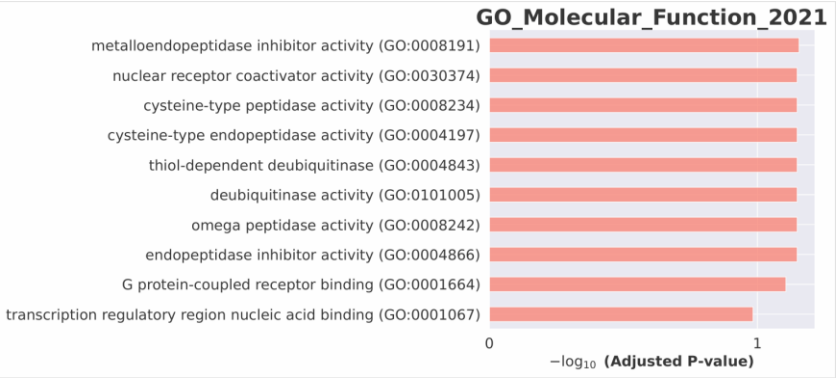

Figure 4. *Gene Ontology (GO) enrichment analysis for the Molecular Function category, identifying the top ten most significantly enriched molecular functions associated with the analysed gene set. The analysis is presented as a horizontal bar plot, where each bar represents an enriched GO term, listed with its corresponding GO accession number (e.g., GO:0008191). The length of the bar corresponds to the statistical significance of the enrichment, measured as the negative logarithm (base 10) of the Adjusted P-value ({-log} _ {10} ({Adjusted P-value})). Terms with a larger {-log} _ {10} ({Adjusted P-value}) are more significantly enriched.*

Gene Ontology (GO) Biological Process Analysis

The Gene Ontology (GO) Biological Process 2021 analysis (Figure 5) specifically focused on cellular behaviour and immune function, revealing that the most significant enriched term was positive regulation of cellular component movement ({-log} _ {10}P \approximately3.0). Furthermore, other highly enriched processes were grouped into key biological categories: Immune/Vascular Interactions, which included multiple terms related to leukocyte adhesion and circulation (such as regulation of leukocyte adhesion to vascular endothelial cell); Haematopoiesis, encompassing terms associated with blood cell development (including positive regulation of myeloid cell differentiation and positive regulation of erythrocyte differentiation); and Cellular Organization, which featured processes like Progressive multifocal leukoencephalopathy (PML) body organization and protein K48-linked deubiquitination.

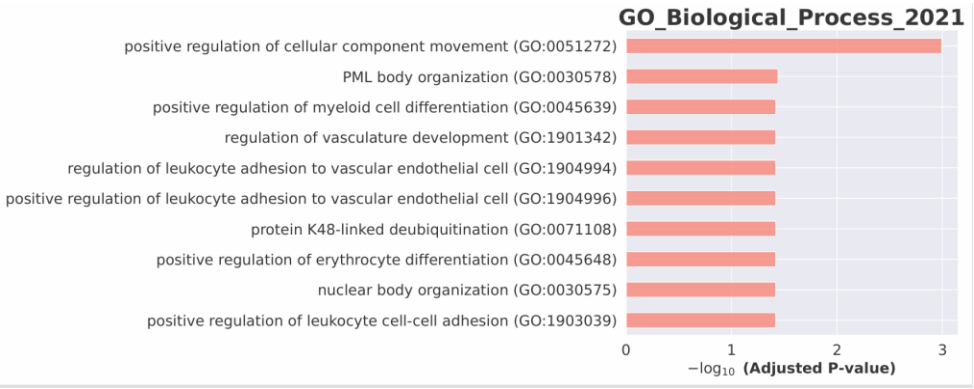

Figure 5. *The figure presents the top ten most significantly enriched Gene Ontology (GO) terms for the Biological Process category, reflecting the primary biological roles of the analysed gene set. The data is displayed as a horizontal bar plot where the statistical significance of each enriched term is represented by the length of its bar, corresponding to the*

negative logarithm (base 10) of the Adjusted P-value ({-log} _ {10} ({Adjusted P-value})). Each bar is labelled with the GO term and its accession number.

scGREAT Consensus GRN and Hub Identification

The scGREAT pipeline successfully generated a Consensus GRN by integrating sparse network data with expression correlation. After filtering to the Top 5% of edges, the network was visualized, focusing on the Top 200 Hubs (Figure 6).

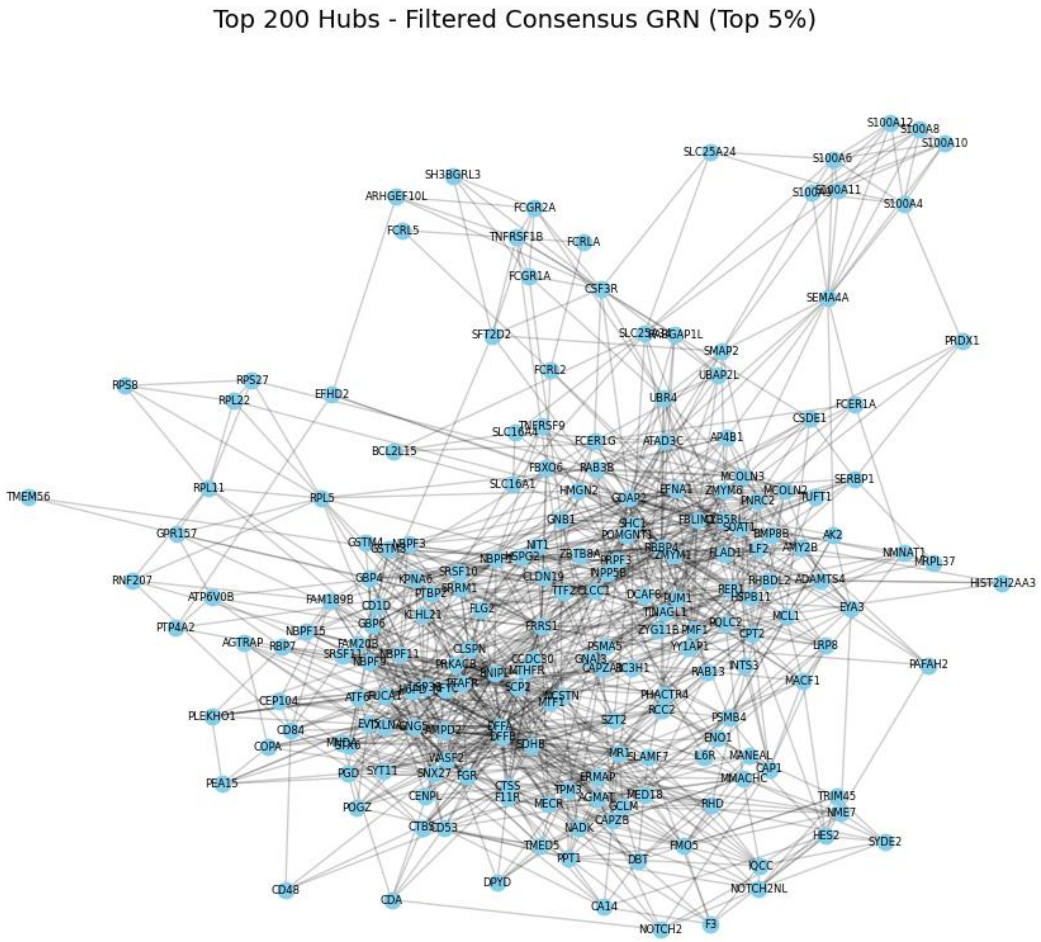

Figure 6. *The figure displays a highly interconnected network representing the Top 200 Hubs identified within a Filtered Consensus Gene Regulatory Network (GRN). The GRN models the regulatory interactions among a set of genes.*

This visualization revealed a complex, highly interconnected structure, particularly a dense central cluster of nodes. The connectivity confirmed that these 200 genes serve as critical regulatory hubs, interacting with numerous other genes and likely acting as central coordinators for the biological processes identified in the enrichment analysis (Section 4.2).

Discussion

The application of scGPT-based clustering successfully identified six discrete gene programs (GPs) whose activation patterns were found to be highly restricted by immune cell lineage, effectively demonstrating the capability of single-cell methodologies to resolve subtle yet crucial transcriptional differences. This specificity is evident in Gene Program 1, which showed high activation in CD14+ Monocytes cells central to acute inflammation and phagocytosis but was entirely absent in Plasmacytoid Dendritic Cells. Conversely, Gene Program 2, prominent in Plasma Cells and CD20+ B Cells, clearly anchors its function within the adaptive humoral immunity compartment. Further validation of this lineage-specific detection is provided by Gene Program 0, which was highly activated exclusively in Erythrocytes, suggesting a non-immune housekeeping or structural role. The visualization of these relationships via the Cosine Similarity Network confirmed the interconnectedness and functional proximity of the identified clusters, suggesting shared or closely related downstream signalling pathways among different programs.

The subsequent functional enrichment analysis, integrating results from Reactome and Gene Ontology (GO), provided a profound biological context for these gene programs, revealing a significant and consistent focus on cellular fate decisions and the Ubiquitin-Proteasome System (UPS). The Reactome analysis highlighted Cellular Senescence and Oncogene Induced Senescence as top enriched pathways, suggesting that the identified gene programs are actively involved in regulating the lifespan and functional exhaustion of immune cells a key mechanism in both aging and chronic inflammatory conditions (Coppé et al., 2010; López-Otín et al., 2023). This theme is intricately linked to protein regulation, as confirmed by highly significant GO Molecular Function terms such as thiol-dependent deubiquitinase and deubiquitinase activity. This strong enrichment of DUB related terms confirms that the post-translational modification and stability of key regulatory proteins are a primary axis of control, a finding critical given the role of DUBs in governing immune signalling pathways like NF-kB activation (Hsu et al., 2024).

Furthermore, the GO Biological Process analysis tied this protein regulation directly to core cellular dynamics and immune function. The most significant term, positive regulation of cellular component movement, alongside others related to leukocyte adhesion to vascular endothelial cell, suggests that the UPS/senescence pathways are crucial for regulating the motility, trafficking, and homing capabilities of immune cells (Lunin et al., 2022). This

synthesis indicates that the identified gene programs do not simply define cell identity but actively govern the mobility and operational readiness of the immune cells.

Finally, the scGREAT Consensus Gene Regulatory Network (GRN) analysis provided the mechanistic framework for these observations. The visualization of the Top 200 Hubs as a complex, highly interconnected network demonstrates that a small subset of genes acts as central coordinators for the entire system. These hubs are likely to include the key transcription factors and DUBs that drive the observed senescence and regulatory phenotypes, providing a robust explanation for how the few identified gene programs can collectively influence such a broad yet specific range of fundamental immune processes. The high connectivity of these hubs makes them high-value targets for future studies seeking to therapeutically modulate immune cell longevity and function. Both Models (scGPT and scGREAT) indicated high performance capacity in gene regulatory network inferences, although scGREAT showed a more detailed network showing the regulatory relationships between the genes.

The highly significant enrichment of pathways related to the Ubiquitin-Proteasome System (UPS) and cellular senescence represents a major finding that provides crucial mechanistic insight into the function of the identified gene programs. The dominance of terms like Ub-specific Processing Proteases and DUB (deubiquitinase) activity is particularly striking, suggesting that the post-translational modification and degradation of proteins rather than transcriptional activation alone are the principal regulatory checkpoints within these immune cells.

The DUBs are known to be the master regulators of numerous critical immune signals. They function by removing ubiquitin tags, thereby stabilizing and activating key transcription factors, adaptor proteins, and receptors that drive immune responses (Hsu et al., 2024). For instance, DUBs regulate the stability of NF-kB components and interferon regulatory factors (IRFs). Therefore, the gene programs identified, particularly the highly connected regulatory hubs, likely contain DUBs that exert a powerful control over the intensity and duration of an immune response. Their localized expression in specific lineages, like Monocytes (Gene Program 1), implies that different immune cell types employ distinct DUB families to fine-tune their unique activation thresholds and cytokine production.

The strong link to Cellular Senescence integrates the identified DUB activity with immune cell longevity and functional fate. Senescence in immune cells known as Immunosenescence

is not merely a state of irreversible growth arrest but is often a programmed response to chronic stimulation or stress, leading to a pro-inflammatory secretory profile (the Senescence-Associated Secretory Phenotype or SASP) (Lunin et al., 2022). The UPS plays a direct role in this process by controlling the degradation of cell cycle inhibitors such as p53 or p21, and the stability of proteins that regulate chromatin structure. Given the functional overlap, it is highly probable that the identified gene programs are involved in regulating the transition of immune cells from an active state to a senescent state. This has profound clinical implications, as dysregulated senescence contributes to age-related decline in vaccine efficacy, autoimmunity, and chronic inflammation (Chen et al., 2022). Therefore, the hubs identified in the GRN could represent novel targets to re-program senescent immune cells or prevent immune exhaustion.

Finally, the convergence of UPS/DUB regulation with the GO Biological Process terms related to cell motility and adhesion provides a compelling functional model. Immune cell trafficking (leukocyte adhesion and circulation) is inherently dynamic and requires rapid, precise changes in protein localization and function (Carman & Martinelli, 2015). The quick, reversible nature of protein ubiquitination makes it the ideal mechanism for controlling these fast-paced processes. The data thus suggest a model where DUB-mediated stabilization of specific motility proteins allows immune cells to quickly adjust their migratory and homing patterns in response to environmental cues, linking the core molecular machinery of protein stability (UPS) directly to the cell's crucial role in surveillance and inflammation resolution.

Methodological Limitations and Future Directions

While the scGPT-based analysis yielded robust and biologically meaningful results, it is essential to acknowledge the inherent methodological limitations that constrain the current interpretation and mandate future validation. The use of Louvain clustering identified six gene programs; however, the initial filtering process (restricting the analysis to 1,200 genes with successfully retrieved embeddings) means that other functionally relevant, but perhaps lower-expressed or more sparsely distributed, gene programs may have been excluded. Furthermore, the Louvain algorithm, while effective, is susceptible to resolution parameters, meaning that slight adjustments could either fragment these six programs into more subtle sub-clusters or merge distinct programs, necessitating a deeper sensitivity analysis to confirm cluster stability.

A more significant constraint lies in the nature of the Gene Regulatory Network (GRN) analysis. The scGREAT pipeline generated a *Consensus* GRN and identified regulatory hubs based on statistical correlation and inferred interactions, which are inherently non-causal. The high connectivity observed in the Top 200 Hubs (Figure 6) suggests their importance, but it does not definitively prove the directionality or direct physical binding of regulatory events. For instance, a hub identified as coordinating senescence could be the result of upstream signalling rather than the initiator of the process. Therefore, the findings presented must be treated as hypotheses-generating. Future experimental validation, using techniques such as chromatin immunoprecipitation (ChIP-seq) or gene perturbation studies targeting the identified hub genes, would be required to establish definitive causal relationships between the regulatory hubs, DUB activity, and immune cell fate.

Conclusion

In conclusion, this project successfully employed a comprehensive computational approach, integrating scGPT-based gene program identification, functional enrichment analysis, and Gene Regulatory Network (GRN) modelling, to characterize the functional landscape of human immune cells. The analysis revealed six distinct, lineage-specific gene programs whose expression patterns cleanly delineate major immune subsets, such as Monocytes and B/Plasma Cells. Critically, the functional annotation provided a unified mechanistic explanation, strongly converging on the regulation of cellular fate through the Ubiquitin-Proteasome System (UPS) and the induction of cellular senescence. The identification of highly interconnected regulatory hubs within the GRN further pinpoints the master coordinators responsible for integrating the observed DUB activity with fundamental immune processes like cell trafficking and longevity. These findings move beyond descriptive characterization to offer mechanistic hypotheses, positioning the identified regulatory hubs and UPS components as novel, high-priority therapeutic targets for modulating immune function in the context of chronic disease, aging, and immunotherapies. This work lays the foundation for targeted experimental validation to translate these computational predictions into functional biological insights.

References

1. Aalto, A., et al. (2020). Gene regulatory network inference from sparsely sampled noisy data. *Nature Communications, 11*(1), 1-13. https://doi.org/10.1038/s41467-020-17217-1

2. Afonja, T., Sheth, I., Binkyte, R., Hanif, W., Ulas, T., Becker, M., & Fritz, M. (2024). LLM4GRN: Discovering causal gene regulatory networks with LLMs Evaluation through synthetic data generation. *arXiv Preprint.* https://doi.org/10.48550/arXiv.2410.15828

3. Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J. C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods, 14*(11), 1083-1086. https://doi.org/10.1038/nmeth.4463

4. Aldridge, S., & Teichmann, S. A. (2020). Single cell transcriptomics comes of age. *Nature Communications, 11*(1), 4307. https://doi.org/10.1038/s41467-020-18158-5

5. Carman, C. V., & Martinelli, R. (2015). T lymphocyte–endothelial interactions: Emerging understanding of trafficking and antigen-specific immunity. *Frontiers in Immunology, 6*, 603. https://doi.org/10.3389/fimmu.2015.00603

6. Chen, J., Deng, J. C., & Goldstein, D. R. (2022). How aging impacts vaccine efficacy: Known molecular and cellular mechanisms and future directions. *Trends in Molecular Medicine, 28*(12), 1100-1111. https://doi.org/10.1016/j.molmed.2022.09.008

7. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., & Wang, B. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods, 21*(8), 1470-1480. https://doi.org/10.1038/s41592-024-02201-0

8. Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Reports*, *2*(3), 666–673. https://doi.org/10.1016/j.celrep.2012.08.003

9. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One, 5*(9). https://doi.org/10.1371/journal.pone.0012776

10. Hsu, S. K., Chou, C. K., Lin, I. L., & Chen, J. K. (2024). Deubiquitinating enzymes: Potential regulators of the tumor microenvironment and implications for immune evasion. *Cell Communication and Signaling, 22*(1), 259. https://doi.org/10.1186/s12964-024-01633-7

11. Islam, S., Kjällquist, A., Moliner, E., Zajac, P., Fan, J. B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1184–1192. https://doi.org/10.1101/gr.110882.110

12. Lunin, S. M., Novoselova, E. G., Glushkova, O. V., Parfenyuk, S. B., Novoselova, T. V., & Khrenov, M. O. (2022). Cell senescence and central regulators of immune response. *International Journal of Molecular Sciences, 23*(8), 4109. https://doi.org/10.3390/ijms23084109

13. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

14. Moerman, T., Aibar, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., & Van Helden, J. (2019). GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics, 35*(12), 2159-2161. https://doi.org/10.1093/bioinformatics/bty916

15. Rossner, T., Li, Z., Balke, J., Salehfard, N., Seifert, T., & Tang, M. (2025). Integrating single-cell foundation models with graph neural networks for drug response prediction. *arXiv Preprint.* https://arxiv.org/abs/2504.14361

16. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., N, W., N, D., J, L., D, T., R, T., B, S., H, C., J, F., B, S., J, W., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

17. Wang, Y., Chen, X., Zheng, Z., Huang, L., Xie, W., Wang, F., Zhang, Z., & Wong, K. C. (2024). scGREAT: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. *iScience, 27*(4), 109352. https://doi.org/10.1016/j.isci.2024.109352

18. Xie, G., Huang, J., Guo, W., Xu, J., Wang, Y., & Wang, Z. (2024). Transformer-based single-cell language model: A survey. *arXiv Preprint.* https://arxiv.org/abs/2407.13205