Final Project Report

Maha Naim

Yin, X., Hu, L., Zhang, Y. *et al.* PDGFB-expressing mesenchymal stem cells improve human hematopoietic stem cell engraftment in immunodeficient mice. *Bone Marrow Transplant* 55, 1029–1040 (2020). https://doi.org/10.1038/s41409-019-0766-z

INTRODUCTION

Bone marrow is home to two types of stem cells: hematopoietic and stromal. The former is responsible for producing blood cells, and the latter is responsible for producing fat, bone, and cartilage. Studies on hematopoietic stem cells (HSCs) are expanding rapidly within the realm of RNA-sequencing research given their essential role in maintaining the homeostasis of HSCs. This biological function is sensitive to disruption by various sources, such as chemotherapeutics, and thus, establishes the importance in better understanding HSCs. Currently, a growing research focus is the lack of appropriately humanized microenvironments to study human immunity, especially within the context of transplantation for treatment of hematological malignancies. Authors, Yin and others, expanded upon this by studying human hematopoietic engraftment using humanized niche cells in their paper *PDGFB-expressing mesenchymal stem cells improve human hematopoietic stem cell engraftment in immunodeficient mice*. In the study, the effects of human factors EGF, FGF2, and PDGFB on MSCs have been compared with respect to HSC hematopoietic regeneration. Human mesenchymal stem cells (MSCs) were transduced by these factors, from which PDGFB-induced MSCs were found to best enhance transplanted human HSC renewal in secondary transplantation. This finding exhibited PDGFB-induced MSCs improved ability to survive and proliferate following transplantation in a humanized cellular environment. Conclusively, the study revealed the efficacy of MSC-mediated factors in human HSC engraftment and offers potential applications in *in vivo* HSC expansion.

For purposes of the project, RNA-sequencing data of cultured GFP-induced MSCs (GFP-MSCs) and PDGFB-induced MSCs (PDGFB-MSCs) was utilized to recreate Figures 4A and 4B. Figure 4A displays a volcano plot and Figure 4B shows a heatmap. To generate these figures, two different analyses were performed on a normalized Transcripts Per Million (TPM) matrix containing the transcription frequencies of approximately 39,037 genes according to 6 different samples where 3 GFP-MSC groups represented the control and the remaining 3 PDGFB-MSC groups were the treatment samples. This data was obtained using the GEO ID referenced in the paper: GSE113857. Upon contact with the author, the statistical data from the study's DESeq analysis was also provided.

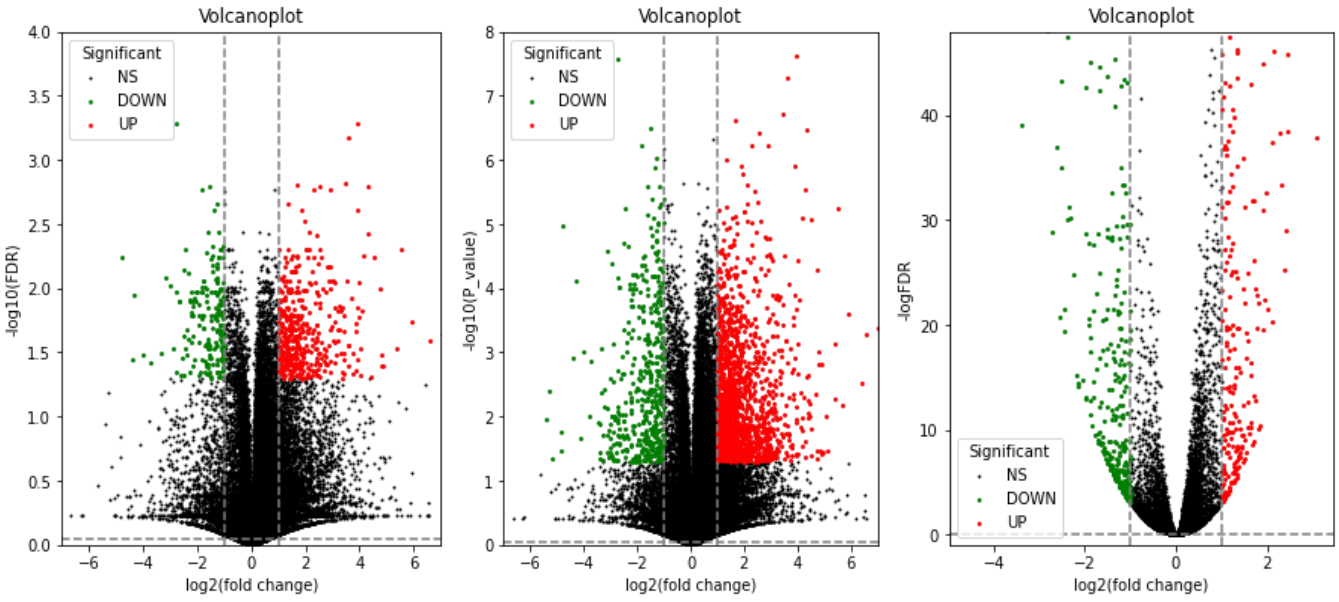
METHODS

To generate the volcano plot from Figure 4A, statistical analysis of the TPM matrix was necessary. The data file was first read in as a Pandas dataframe and indexed according to the gene symbol. As a sanity check, no missing values were found in the dataframe. Averages across each respective experimental group were then calculated using Pandas. For example, for each gene symbol, 3 row values were averaged according to columns that were denoted by either "GFP-MSC" or "PDGFB-MSC." Using these averages, fold change values were calculated by dividing the mean value for the GFP-MSC group by that of the PDGFB-MSC group. To preserve the quality of the data, infinite values were replaced with "NaN" and these missing values were then filled in with a value of zero. Using the fold change values, the log2 fold change (log2FC) was computed using Numpy. Additionally, similar to the method applied with the average

calculations, standard deviations across each respective experimental group were calculated using Pandas. Using the Statsmodels module, p-values were calculated for each experimental group and gene, and then corrected for by the module's false discovery rate (FDR) correction procedure. Three different plots were generated using Matplotlib: the first two used the calculated statistical measures and the final plot used the statistical measures reported in the DESeq2 analysis file provided by the author. The first plot visualized log2FC against the computed FDR values, the second plot visualized log2FC against the computed p-values, and the third plot visualized log2FC against FDR as computed by the authors during their DESeq2 analysis. A log2FC threshold of ≥ 2 and FDR threshold of ≤ 0.05 were represented in the plots.

To generate the heatmap from Figure 4B, a distance calculation and clustering algorithm were implemented on the TPM matrix. The data was read in as a Pandas dataframe and subsetted so that only the 23 genes reported in the paper's heatmap were retained from the TPM data. The dataset was then indexed by the gene symbol. To compute the distance matrix, a number of different distance or correlation methods were attempted including: Pearson's, Spearman's, and Euclidean distances. After transposing the dataframe, the best fitting distance calculation was offered by the Pearson correlation method. The optimal clusters were determined according to various linkage methods including: single, complete, and Ward. However, implementing the Ward clustering algorithm on the computed Pearson distances using the SciPy Hierarchical Clustering Linkage module best replicated the paper's results. To plot, Seaborn was used to generate a clustered heatmap on the subsetted data according to the computed clusters.

FIGURES AND CONCLUSIONS



(A) **(B)** **(C)**

Figure 1. Volcano plot of PDGFB-MSCs versus GFP-MSCs mRNA profiling showing up-regulated (red) and down-regulated (green) genes. Reproducibility attempt of Figure 4A in the paper. (A) Plot generated according to FDR values and log2FC with FDR ≤ 0.05, $|\log2FC| \geq 2$. **(B)** Plot generated according to FDR and log2FC computed with p-value ≤ 0.05,

|log2FC| ≥ 2. **(C)** Plot generated according to FDR and log2FC reported in the author's DESeq2 output with FDR ≤ 0.05, |log2FC| ≥ 2.

The volcano plot shown in Figure 4A of the paper aimed to provide insight into the global mRNA profiling of up-regulated and down-regulated genes in PDGFB-MSCs and GFP-MSCs. This plot features the -log10FDR values according to log2FC. By doing so, information on the molecular agents involved in promoting human engraftment became available. In an attempt to replicate this figure, three total plots were produced as shown in Figure 1. Like the paper, Figure 1A plots the log2FC and FDR values that were computed during the analysis workflow. Though the overall shape of the scatterplot vaguely resembles that of Figure 4A, the FDR scale varies significantly between both plots. To achieve a more quantitative comparison, the amount of up-regulated and down-regulated genes were identified and shown to be higher by 19 and 90, respectively, for Figure 1A than in Figure 4A of the paper. Similarly, Figure 1B demonstrates an even higher discordance in gene count, as it plots the p-values computed during the analysis workflow rather than the FDR values like in Figure 1A.

As a sanity check, Figure 1C was generated using the data from the DESeq2 analysis that was performed in the study. Expectedly, the number of up-regulated and down-regulated genes, 250 and 350, respectively, matched those reported by the paper. The resulting plot almost exactly replicates that of the paper, given that the plotted values were computed according to the outlined DESeq2 procedure in the study. This check offered some confidence in the accuracy of the plotting methodology that was implemented for Figures 1A and 1B and indicated that any discrepancies between these plots and Figure 1C must be related to the statistical computing methodology. The difference in the scale of log2FC was minimal between all three plots, therefore, the main area of concern is likely the p-value calculation and FDR correction. The current p-values were computed using a t-test and may have differed from the computation method that DESeq2 applies. For future attempts, a different p-value statistical test, and possibly an alternative FDR correction procedure, should be considered.

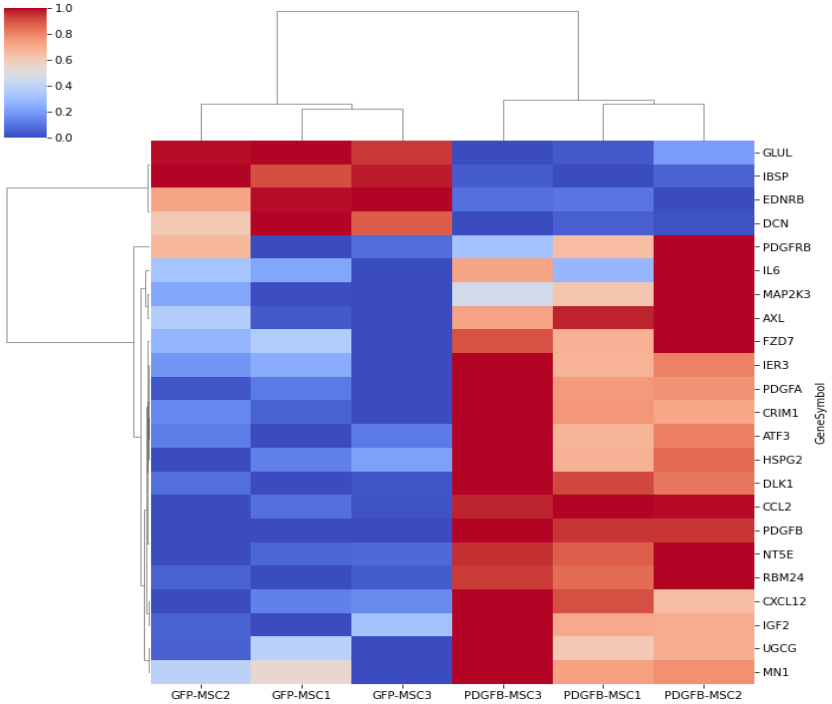


Figure 2. Heatmap of PDGFB-MSCs versus GFP-MSCs differentially expressed genes showing up-regulated (red) and down-regulated (blue) genes according to their respective sample group. Reproducibility attempt of Figure 4B in the paper.

The heatmap in Figure 4B was produced to evaluate the functional relationships between the MSC autocrine/paracrine genes, stemness-related genes, and differentiation genes that are differentially expressed between PDGFB-MSCs and GFP-MSCs. Upon analysis of the heatmap generated by the Pearson correlation and Ward's cluster variance method, several of the conclusions reached by the study's Figure 4B are also supported by Figure 2. For example, Yin et al. report, "Genes indicative of MSC stemness (NT5E, RBM24, CRIM1, UGCG, MN1) [29] were upregulated and differentiation genes (IBSP, EDNRB) [29] were downregulated in PDGFB-MSCs compared to GFP-MSCs." These findings, as well as others, are consistent with the heatmap results produced in Figure 2. Given the concordance in results between Figure 2 and Figure 4B, there is confidence in the distance and cluster methodology implemented during the analysis to generate the clustered heatmap. Furthermore, since this heatmap is clustered, additional interpretation on the close-relatedness of the samples and genes is available that is not currently offered by Figure 4B. For example, the dendrogram along the x-axis clearly indicates two unique clusters that are representative of the PDGFB-MSC group and the GFP-MSC group. With regards to the dendrogram on the y-axis, approximately six unique clusters are indicated among the 23 genes of interest. An interesting follow-up analysis could be to explore the significance of these gene clusters as they relate to MSC functionality.

To conclude, PDGFB has shown to promote *ex vivo* expansion of MSCs longitudinally and is a promising clinical target in immunotherapeutics. The combined use of other factors, along with PDGFB, may be a promising future direction to continue enhancing the niche-supporting effects of MSCs.

ATTACHED FILES

1. Python Code - Naim_Maha_Project.ipynb
2. Report - Final_Project_Report.pdf
3. TPM Dataset - GSE113857_TPM_matrix.txt
4. DESeq2 Output - DEGs_tablefinalone.txt
5. Figures - Plot1A_Volcano_Plot_FDR.png, Plot1B_Volcano_Plot_Pvalue.png, Plot1C_Volcano_Plot_authorFDR.png, Plot2_Clustered_Heatmap.png

Works Cited

1. Yin, X., Hu, L., Zhang, Y. *et al.* PDGFB-expressing mesenchymal stem cells improve human hematopoietic stem cell engraftment in immunodeficient mice. *Bone Marrow Transplant* 55, 1029–1040 (2020). https://doi.org/10.1038/s41409-019-0766-z