Gene Length Correlates with Age-related Gene Regulation in Human Hematopoietic Stem and Progenitor Cells that Predispose to Myeloid Leukemia

Hannah Thomas
Nashua High School South, Nashua NH

Introduction

The human aging process is associated with a decline in the function of hematopoietic stem and progenitor cells (HSPCs) and an increasing susceptibility to myeloid malignancies. Adelman et al. studied the epigenomic and transcriptomic alterations in HSPCs during aging. They gathered bone marrow mononuclear cells (MNCs) from young and aged donors without a history of hematologic cancer and used diverse sequencing techniques for data collection and analysis. Their research produced several new findings. Notably, they discovered that differentially methylated regions (DMRs) observed in aged HSPCs were also present in both young and elderly patients with acute myeloid leukemia (AML), indicating that differential gene expression in aged HSPCs may cause a potential predisposition to myeloid malignancies. Furthermore, they identified that the epigenomic changes stemmed from HSPC reprogramming rather than existing factors and were unique to HSPCs. In this study, we explored further the mechanisms driving these changes in aged HSPCs and their connections to cancer research.

In particular, we studied the specific changes in the gene expression of HSPCs by analyzing data presented by Adelman et al. [1]. The dataset includes all differentially expressed genes in HSPCs along with their expression fold change, which indicates if the gene is upregulated or downregulated. We analyzed if there is a statistically significant correlation between the change in regulation and the main architectural parameters of the gene: gene length, guanine-cytosine (GC) content, count of transcripts, and count of exons. Data for these parameters are available for download in the Ensemble111 BioMart database. We hypothesized

that these gene architectural parameters correlate with age-related gene regulation in human hematopoietic stem and progenitor cells that may predispose to myeloid malignancies.

Method

First, our investigation needed to determine what specific changes were occurring in the gene expression of HSPCs. For this, we obtained Table S6 from the Supplementary data of Adelman et al., which included the IDs of all differentially expressed genes in HSPCs, as well as other information, like the base two logarithm of their fold change [1]. Additionally, we created data files to compute the values of the aforementioned architectural parameters for each differentially expressed gene. The gene parameters along with the gene stable ID for every gene in the human genome are available for download in the Ensembl BioMart database. We curated three data files to use in our analysis. The first file had basic features that included the starting and ending base pairs along with the transcript stable IDs of each transcript from all genes. This file would provide the necessary information to calculate the length and count of transcripts for any gene. The second file contained the GC content listed as a percentage to use in the program. The third file listed the IDs of each gene's exons to compute the count of exons per gene.

We wrote a Python program for a systematic analysis of the change in gene expression with its architectural parameters. We initially imported the scipy, numpy, and matplotlib libraries and created Genes and DiffExpGenes classes. The Genes class would store the name, type, starting base pair, ending base pair, length, count of transcripts, count of exons, and GC content of every gene. The DiffExpGenes class would store the data from Table S6 as well as the length, GC content, transcript count, and exon count of all the genes listed in that file. We later separated the differentially expressed genes into two arrays: one with all the upregulated genes and one with all the downregulated genes. To efficiently store and access the data, we used a dictionary

data structure called GeneDict with all the differentially expressed genes and their respective attribute values from the Genes class.

To find gene length and transcript count, we read the transcript file while keeping track of the smallest starting base pair and the largest ending base pair of each differentially expressed gene. Gene length was calculated by subtracting the smallest starting base pair from the largest ending base pair. The count of transcripts was calculated by tallying the instances where transcripts shared the same gene ID. We read the GC content file and populated the GC Content attribute for each gene in GeneDict. Similarly, we read the exon file and calculated the exon count for all genes in the gene dictionary.

We used the data populated in GeneDict to construct eight arrays to run statistical tests. Each parameter was represented by two arrays: one with the upregulated genes' values and the other with the downregulated genes' values. We used various statistical tests to compare the two arrays for each gene architectural parameter and determine if there was a significant correlation between that parameter and gene regulation, specifically, the t-test, the Kolmogorov-Smirnov (K-S) one-sample and two-sample tests, and the Mann-Whitney U test. We ran these tests and generated box plots to display the data using functions from the installed Python libraries.

Results

We hypothesized that the gene architectural parameters, gene length, GC content, transcript count, and exon count correspond to age-related epigenetic modifications in HSPCs that could predispose to myeloid malignancies. The t-test results showed that none of the parameters have a statistically significant correlation to gene regulation ($p\text{-value} > 0.05$).

T-Test Results

Gene Architecture Parameter	Statistic	p-value
Gene Length	1.28	0.20
GC Content	-1.45	0.15
Exon Count	1.44	0.15
Transcript Count	-0.09	0.93

The Mann-Whitney U test results showed a strong, statistically significant correlation between gene length and gene regulation (p-value < 0.05). However, none of the other parameters showed a statistically significant correlation (p-value > 0.05).

Mann-Whitney U Test Results

Gene Architecture Parameter	Statistic	p-value
Gene Length	169909	0.0003
GC Content	146135.5	0.37
Exon Count	157947	0.18
Transcript Count	148635	0.68

The Kolmogorov-Smirnov two-sample test provided the same result as the Mann-Whitney U test: that only gene length correlates with gene regulation.

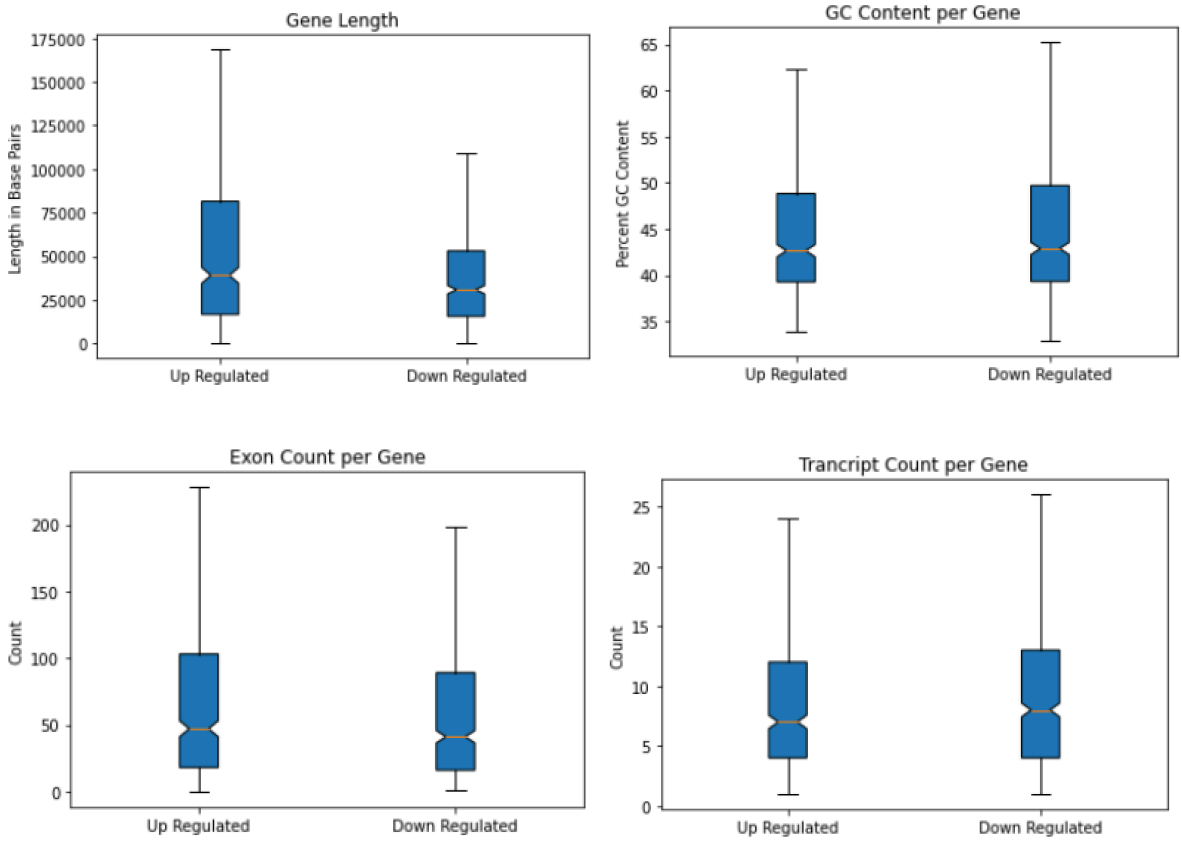
Kolmogorov-Smirnov Two-sample Test Results

Gene Architecture Parameter	Statistic	p-value
Gene Length	0.16	1.83E-06
GC Content	0.06	0.31
Exon Count	0.06	0.32
Transcript Count	0.05	0.58

A significant difference between the t-test and the other two statistical tests is that the t-test assumes the sample data follows a normal distribution whereas the others do not. If the parameter values are not normally distributed, the t-test is not reliable. We used the

Kolmogorov-Smirnov one-sample test to check for normality of the data and the p-values of those tests showed that none of the architectural parameters are normally distributed, confirming that the t-test results are not reliable. Therefore, we concluded that gene length has a strong, statistically significant correlation to epigenetic changes in aged HSPCs.

The figures below display box plots of the architectural parameters of upregulated and downregulated genes.



This phenomenon may be explained by the accumulation of transcriptional stress with age. Gyenis et al. state that stress caused by DNA damage hinders the ability of RNA polymerase II to function effectively during transcription [2]. This would lead to a transcriptome imbalance dependent on gene length, where the longer the gene, the more likely it is to acquire stress.

Conclusion

This research provides a detailed explanation of the relationship between gene length and developments in age-related regulation in HSPCs. Stoeger et al. corroborate our conclusion with their study on transcriptional changes in general mice and human cells [5]. The use of rigorous statistical tests ensures the reliability of our findings, creating a more robust understanding of the processes that occur during the aging of human genes in HSPCs. To our knowledge, no similar studies of human HSPCs have been reported in the literature.

The significance of our finding is that these length-related changes in gene regulation may predispose to myeloid malignancies as older people are more susceptible to developing illnesses such as acute myeloid leukemia (AML), and prior research supports this claim. As these expression changes are known to predispose to leukemia, scientists can target to modify the expression of genes of specific architecture for the prevention of leukemia. Furthermore, this information can be used as a biomarker to evaluate the risk of developing leukemia.

Bibliography

1. Adelman, E. (2019, May 13). *Aging human hematopoietic stem cells manifest profound epigenetic reprogramming of enhancers that may predispose to leukemia*. Cancer discovery. https://pubmed.ncbi.nlm.nih.gov/31085557/
2. Gyenis, A. (2023, January 19). *Genome-wide RNA polymerase stalling shapes the transcriptome during aging*. Nature genetics. https://pubmed.ncbi.nlm.nih.gov/36658433/
3. IBM Corporation. (2021, March 22). *One-sample Kolmogorov-Smirnov Test*. IBM. https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=tests-one-sample-kolmogorov-smirnov-test

4. RTC Lab. (n.d.). *Kolmogorov Smirnov Two Sample Test*. FinMath. https://rtmath.net/assets/docs/finmath/html/d52f0b5b-0fc2-448f-84e1-ce311d904f01.htm
5. Stoeger, T. (2022, December 9). *Aging is associated with a systemic length-associated transcriptome imbalance*. Nature aging. https://pubmed.ncbi.nlm.nih.gov/37118543/
6. *T-test, Chi-square, ANOVA, regression, correlation...* Datatab. (n.d.). https://datatab.net/tutorial/mann-whitney-u-test
7. *T-test, Chi-square, ANOVA, regression, correlation...* Datatab. (n.d.-a). https://datatab.net/tutorial/t-test
8. Wadhwa, R. R. (2023, January 16). *T test*. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK553048/