Title

Emine Guven and Sevinc Akcay

Address

**Abstract:** Aging is defined as an increase in failure (mortality) rate that is irreversible and in biological mechanisms leading to progressive functional decline and increased risk for disease and death. We characterized aging at the gene expression level using GSE32719 data set publicly available at gene expression omnibus (GEO) and ArrayExpress. Using Biobase, GEOquery, gplots packages out of which 96 genes were chosen according to criteria of p-value < 0.05 and fold change < 0.05 for further comparative analysis. We then investigate gene ontology (GO) analysis for these genes. Subsequently, the DEGs were subjected to biological process and functional analysis which deciphers the increase in hematopoietic stem cell population and functional decline. More importantly, the results reveal FLT3 and SPP1 genes which are also known as controlling many important cellular processes such as the growth and division (proliferation) and survival of cells, particularly of early blood cells called hematopoietic progenitor cells and T-lymphocyte activation (ETA-1).

**Keywords:**

**Reference** to this paper should be made as follows: Author. (xxxx) ‘Title’, *Int. J. xxxxxxxxxxx xxxxxxxxxxx*,

**Biographical notes:**

*This paper is a revised and expanded version of a paper entitled* ***[title]*** *presented at [****name, location and date of conference****].*

1 Introduction

The process of aging is characterized by a degeneration in the maintenance of homeostatic processes with advancing time, leading to functional decline in a variety of organ and tissue systems and increased risk for many diseases (cardiovascular, several types of cancer, metabolic and neurological diseases) and ultimately death [1]. Aging decreases an organism’s ability to handle environmental and physiological disturbance which also another cause that increases the vulnerability to death [2]. It is important to identify and characterize the genetic and environmental factors (smoking, environmental pollution) that modulate longevity in order to understand the basic mechanism of aging. The complex aging process and wide variability between individuals limited the identification of factors affecting aging [2] . To date, several theories have been proposed which provided beneficial insights to understand the physiological changes during the aging process. The notable ones are immunologic, inflammation, free radical and mitochondrial. To date several hundred genes have been found to associated with aging. Bone marrow is a special site where blood cells are covered with structural stromal cells. It is a spongy adipose tissue found in the bones such as femurs, rib cage, ribs, pelvis and human skull.

Bone marrow is fed with specialized blood vessels and contributes to circulation. The specialized phenestra capillary, called sinusoid, penetrates the extracellular / extracellular matrix (ECM), and ECMs are sponge-like matrices produced by reticular fibroblasts [3].

These proteins and cells place the hematopoietic cells in separate compartments. Similarly, hematopoietic cells, bidothelial cells, stromal cells, ECM, cytokines, growth factors, and chemokines contain special microenvironment in the bone marrow. Bone marrow microenvironment plays an important role in the development and progression of leukemia and other types of cancer [3, 4] .

Recent advances include new markers for HSCs and niche stem cells, systematic analysis of expression patterns of niche factors, genetic tools for functional in vivo identification of niche cells, and improved imaging techniques. Stem cells have the ability to divide for a long time in the living body, to be able to regenerate and transform into other tissue cells by differentiating according to the needs of the body [4, 5]. Various criteria have been determined to define a cell as a stem cell. For example, stem cells can contribute to undifferentiated lines even in vivo without tissue damage. Understanding of the properties of hematopoietic cells was made 40 years ago by a series of seminal experiments demonstrating the ability to form macroscopic colonies on a subset of cells on the bone marrow [6, 7]. Hematopeotic Stem Cells (HSCs) are specific stem cells that are prospectively isolated into tissue and are also the only stem cells in routine clinical use (add citations). Grafts commonly used in the treatment of various diseases such as leukemia and autoimmune cause this condition. A detailed examination of the cellular and molecular properties of HSCs has made studies in clinical use, stem cell identification and use highly effective. Bone marrow microenvironment is an ideal place to support healthy and malignant hematopoiesis [1, 8, 9]

In acute myeloid leukemia (AML), blast cells, which deteriorate during normal maturation, begin to accumulate in the blood together with the bone marrow. The body remains vulnerable because white blood cells cannot form. Erythrocyte and platelet production are disrupted in the bone marrow due to abnormal proliferation of myeloblasts. As a result, anemia infection and platelet count decrease.

Bleeding in AML occurs as the first onset symptoms. AML is not limited to a particular part of the body since its onset, but can spread to the blood, lymphatic tissue and all other organ systems from the bone marrow. As with many other leukemia diseases, it is defined as a malignant systemic disease.

Previous studies addressing age-related changes in human HSC, due to the indirect evaluation of root and progenitor populations, need to support experimental studies with numerical analysis and statistical methods in addition to previous HSC studies in mice [6, 8]. In addition to the clinical results obtained in Pang et al. 2012, gene expressions were analyzed by various and different tools and methods. To describe the characteristics of the elderly human hematopoietic system, which may be prone to age-related hematopoietic function disorders, from healthy, hematologically young, middle, and old human bone marrow samples [10]. HSC and other hemotopoietic progenitor populations are evaluated and it was found that the elderly HSC increased frequently. Different than previous research, here we take into consideration general statistics of micro-array gene expression data globally and locally. Although methods to correct for multiple comparisons have been available for a long time [3-6] (eg, Bonferroni [15]  correction), many of these methods are not well suited for the analysis of microarray data. Mostly because most techniques assume variable independence and many are considered to be stringent quality assurance procedures.

MATERIALS AND METHODS

2.1. Microarray data and preprocessing

Expression data from human bone marrow hematopoietic stem cells were downloaded from the gene expression omnibus (GEO) database with GSE32719 was used [10]. Genomic information ranging from gene sequences to protein structure predictions were obtained. As described by Pang et al [10], these data sets contain a total of 50,000 gene expression of healthy human bone marrow hematopoietic stem cells in groups of 14 young (20–31 years), 5 middle age (42–61), 8 old (65–85) groups. The GSE32719 data set is analyzed by using the GEOquery package [16] in Bioconductor [17] following standard procedures in R studio [18]. The other packages we used in R studio are as the following; Biobase and gplots packages [19].

2.2. Experimental data and analysis codes

Analysis were conducted in the R statistical environment. Sample codes and analysis of GSE32719 data can be found at <https://github.com/emineguven/modelComparison2018>. We separate samples into three conditions provided that young-old , young-middle aged , and middle-old aged.  The data set was normalized by computing the means of the samples of each condition in R programming language. The process on separated samples which is grouped by conditions of young-old, young-middle, and middle old was performed as computing fold-change (biological significance) difference between the means of the conditions. We then find statistical significance using student’s t-test by taking fold cut-off value 0.075 young-old ; fold cut-off value 0.05 young-middle and middle-old and p-value cutoff value 0.01 for each conditions respectively. The genes are screened based on satisfaction of both fold and p-value filtering criteria. We highlight the significantly up-regulated and down-regulated differential expressed genes (DEGs).

2.3. Annotation of differentially expressed genes

Expression measurements annotations for up-regulated and down regulated DEGs for each conditions probes mapped to gene names using Ensemble Biomart server [20]. We choose the database Ensemble Genes 100, filtered by genes and uniport id with the related Affy ids. All characterized genes were carefully investigated and additional components like the Universal Protein resource, and physical properties Gene Ontology (GO) and annotation types were derived using ***G****ene****O****ntology En****ri****chment ana****l****ysis and Visua****l****iz****a****tion Tool* (GOrilla)and DAVID [21–23].

RESULTS

3.1. Experimental data analysis

Microarray data repository are publicly available and an extensive search of these repositories revealed that there is a large number of datasets available involving human aging and from several human organs. Goals of these studies were diverse and included different test conditions, tissue types, and subjects. Analysis of statistical and functional annotation of microarray data can be done using several methods . The data sets used in this study were obtained from public microarray data repositories, i.e., NCBI GEO and ArrayExpress where microarray data are available and associated with aging.

With gene expression result of the GSE32719 data set, we detect Differentially Expressed Genes (**DEGs**) in total 96 genes between conditions from young-old and young-middle aged samples. Up and down regulated gene expressions between **young-old, young-middle**, and **middle-old** conditions shown in Table 1. Here, we detected 24 differentially expressed genes of down regulation, whereas we find 72 up-regulated genes between high and low conditions. We find the up-regulated and down-regulated differentially expressed genes between conditions.

|  |  |  |
| --- | --- | --- |
| Conditions Compared | down-regulated | up-regulated |
| young-old | 9 | 21 |
| young-middle | 15 | 51 |
| middle-old | - | - |

**Table:** Summary of differentially expressed genes by conditions.

However, we could not affirm if the data sets belonged to any specific gender, i.e., male or females. On the other hand, the reason we fail to detect any DEGs in middle-old condition is that the probe IDs of middle-old condition does not map to any of the gene names using Ensemble Biomart server [20].

3.2 Gene Ontology Enrichment Analysis

After successfully detecting the differentially over (up) and under (down) expressed genes for the young-old and young-middle conditions, **GOrilla** is used to identify and visualize enriched GO terms in ranked lists of genes. We do the enrichment analyses for genes in the up and down-regulated results. We have used approximately 21,000 genes as a background gene reference with two-rank listed analyses in GOrilla.

**Gorilla has calculated the 'P-value'** which is the enrichment p-value computed according to the mHG or HG model. This p-value is not corrected for multiple testing of 14548 GO terms.  
  
**The calculated q-value is called 'FDR q-value'** which is the correction of the above p-value for multiple testing using the Benjamini and Hochberg method [24].

For the ith term (ranked according to p-value) the FDR q-value is,

**The value for Enrichment (N, B, n, b)** is defined as follows:

N - is the total number of genes  
B - is the total number of genes associated with a specific GO term  
n - is the number of genes in the top of the user's input list or in the target set when appropriate  
b - is the number of genes in the intersection

**Genes:** For each GO term you can see the list of associated genes that appear in the optimal top of the list. Each gene name is specified by gene symbol followed by a short description of the genes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **GO Term** | **Description** | **P-value** | **FDR q-value** | **Enrichment** | **N** | **B** | **n** | **b** | **Genes** |
| GO:0048545 | response to steroid hormone | 5.51E-05 | 2.92E-01 | 16.3 | 880 | 12 | 18 | 4 | LOX , CASP9 , FLT3 , SPP1 |
| GO:0014070 | response to organic cyclic compound | 6.60E-05 | 1.75E-01 | 10.19 | 880 | 24 | 18 | 5 | LOX , CASP9 , ADA, FLT3 , SPP1 |
| GO:1901361 | organic cyclic compound catabolic process | 3.15E-04 | 5.57E-01 | 10.86 | 880 | 18 | 18 | 4 | AGO3 , FAH , ADA , SPP1 |
| GO:1901575 | organic substance catabolic process | 3.43E-04 | 4.54E-01 | 5.64 | 880 | 52 | 18 | 6 | AGO3 , FUT2 , FAH , ADA , GLA , SPP1 |
| GO:0071407 | cellular response to organic cyclic compound | 7.90E-04 | 8.39E-01 | 14.67 | 880 | 10 | 18 | 3 | CASP9 ,FLT3 , SPP1 |
| GO:0044282 | small molecule catabolic process | 7.90E-04 | 6.99E-01 | 14.67 | 880 | 10 | 18 | 3 | FUT2 , FAH , ADA |
| GO:0042127 | regulation of cell proliferation | 9.15E-04 | 6.93E-01 | 4.73 | 880 | 62 | 18 | 6 | AGO3 , FUT2 , HPGDS , ADA , FLT3 , NMB |
| GO:0009056 | catabolic process | 9.15E-04 | 6.07E-01 | 4.73 | 880 | 62 | 18 | 6 | AGO3 , FUT2, FAH , ADA , GLA , SPP1 |

|  |
| --- |
| Table 1: Summary of GO pathways of all up-regulated genes (in total 21 genes) from the conditions between young and old aged samples. |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **GO Term** | **Description** | **P-value** | **FDR q-value** | **Enrichment** | **N** | **B** | **n** | **b** | **Genes** |
| GO:0032496 | response to lipopolysaccharide | 8.86E-05 | 4.60E-01 | 29.03 | 871 | 10 | 9 | 3 | FOS, KMO, SELP |
| GO:0002237 | response to molecule of bacterial origin | 8.86E-05 | 2.30E-01 | 29.03 | 871 | 10 | 9 | 3 | FOS, KMO, SELP |
| GO:0034654 | nucleobase-containing compound biosynthetic process | 5.95E-04 | 1.00E+00 | 8.8 | 871 | 44 | 9 | 4 | FOS, NPR2, TOPORS, KMO |
| GO:0018130 | heterocycle biosynthetic process | 6.50E-04 | 8.45E-01 | 8.6 | 871 | 45 | 9 | 4 | FOS, NPR2, TOPORS, KMO |
| GO:0019438 | aromatic compound biosynthetic process | 6.50E-04 | 6.76E-01 | 8.6 | 871 | 45 | 9 | 4 | FOS, NPR2, TOPORS, KMO |
| GO:0044271 | cellular nitrogen compound biosynthetic process | 7.09E-04 | 6.14E-01 | 8.42 | 871 | 46 | 9 | 4 | FOS, NPR2, TOPORS, KMO |
| GO:1901362 | organic cyclic compound biosynthetic process | 8.37E-04 | 6.22E-01 | 8.06 | 871 | 48 | 9 | 4 | FOS, NPR2, TOPORS, KMO |

Table 2: Summary of GO pathways of all down-regulated genes (in total 9 genes) from the conditions between young and old aged samples.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **GO Term** | **Description** | **P-value** | **FDR q-value** | **Enrichment** | **N** | **B** | **n** | **b** | **Genes** |
| GO:0006413 | translational initiation | 8.10E-05 | 4.38E-01 | 22.55 | 902 | 3 | 40 | 3 | EIF3C, EIF3CL, EIF3G |
| GO:0001755 | neural crest cell migration | 8.10E-05 | 2.19E-01 | 22.55 | 902 | 3 | 40 | 3 | SEMA4A, PITX2, SHH |
| GO:0001732 | formation of cytoplasmic translation initiation complex | 8.10E-05 | 1.46E-01 | 22.55 | 902 | 3 | 40 | 3 | EIF3C, EIF3CL, EIF3G |
| GO:0022618 | ribonucleoprotein complex assembly | 2.05E-04 | 2.77E-01 | 11.28 | 902 | 8 | 40 | 4 | EIF3C, EIF3CL, SRPK3, EIF3G |
| GO:0071826 | ribonucleoprotein complex subunit organization | 3.57E-04 | 3.86E-01 | 10.02 | 902 | 9 | 40 | 4 | EIF3C, EIF3CL, SRPK3, EIF3G |
| GO:0034622 | cellular protein-containing complex assembly | 4.92E-04 | 4.43E-01 | 5.41 | 902 | 25 | 40 | 6 | EIF3C, NDUFB11, EIF3CL, DIAPH1, SRPK3, EIF3G |
| GO:0090185 | negative regulation of kidney development | 7.61E-04 | 5.87E-01 | 13.53 | 902 | 5 | 40 | 3 | SHH, STAT1, TACSTD2 |
| GO:0071407 | cellular response to organic cyclic compound | 8.76E-04 | 5.92E-01 | 8.2 | 902 | 11 | 40 | 4 | RALB, EL1, DIAPH1, STAT1 |

Table 3: Summary of GO pathways of all up-regulated genes (in total 51 genes) from the conditions between young and middle aged samples.

3.3. GO pathways results of biological processes of DEGs from the conditions between young and middle aged samples

|  |  |
| --- | --- |
| negative regulation of transcription from RNA polymerase II promoter |  |

Figure 1: A ReViGo scatter plot of GO analyses in biological process from (a) down-regulated and (b) up-regulated DEGs in the young-middle conditions. Related GO terms are clustered and presented with bubbles with similar color shades. Bubble colors represent p-values and bubble sizes indicate the relative frequency of the GO terms.

The most and most significant GO pathway results are obtained from the down-regulated DEGs between young and middle aged samples. Figure 1 demonstrates the most important pathways that play role in HSCs between young and middle aged samples. We used REVIGO [25] that summarizes and visualizes long list of gene ontology terms using embedded R-script. Among 44 ontology terms, using REVIGO tool that number decreased to 27 GO terms . Visualizing in the scatter plot, 4 of the biological process terms seem to be the most important ones; negative regulation of transcription from RNA polymerase II promoter, rhythmic process, fat cell differentiation, cellular response to organic cyclic compound, and response to chemical. Similar to down-regulated DEGs, up-regulated DEGs between young and middle aged samples (Fig.1, right) can be summarized as the following; translational initiation, cellular response to organic cyclic compound, and neural crest cell migration.

REFERENCES

[1] M. Fraga and M. Esteller, “Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. Trends Genet. 23, 413-418,” *Trends Genet.*, vol. 23, pp. 413–418, Sep. 2007, doi: 10.1016/j.tig.2007.05.008.

[2] M. P. Rodero *et al.*, “Immune surveillance of the lung by migrating tissue monocytes,” *Elife*, vol. 4, pp. e07847–e07847, Jul. 2015, doi: 10.7554/eLife.07847.

[3] A. KARADAĞ, B. ALTINOK, T. ÖZKAN, and Y. HEKMATSHOAR, “Kemik İliği Stroması: Hücreleri ve Mikroçevresi.”

[4] J. M. Perry and L. Li, “Disrupting the stem cell niche: good seeds in bad soil,” *Cell*, vol. 129, no. 6, pp. 1045–1047, 2007.

[5] E. Karaöz, “Ovalı E,” *Kök hücreler. Derya Kitabevi, Trabzon*, 2004.

[6] D. J. Rossi *et al.*, “Cell intrinsic alterations underlie hematopoietic stem cell aging,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 26, pp. 9194–9199, 2005, doi: 10.1073/pnas.0503280102.

[7] D. Bryder, D. J. Rossi, and I. L. Weissman, “Hematopoietic stem cells: The paradigmatic tissue-specific stem cell,” *Am. J. Pathol.*, vol. 169, no. 2, pp. 338–346, 2006, doi: 10.2353/ajpath.2006.060312.

[8] A. S. Çakmak, “Biyofiziksel ve Biyokimyasal Uyaranlarla Desteklenmiş Doku İskeleleri ile Mezenkimal Kök Hücrelerin Osteojenik Farklılaşmasının İncelenmesi,” 2014.

[9] K. Kuranda *et al.*, “Age-related changes in human hematopoietic stem/progenitor cells,” *Aging Cell*, vol. 10, no. 3, pp. 542–546, 2011, doi: 10.1111/j.1474-9726.2011.00675.x.

[10] W. W. Pang *et al.*, “Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 50, pp. 20012–20017, 2011, doi: 10.1073/pnas.1116110108.

[11] A. L. Tarca, R. Romero, and S. Draghici, “Analysis of microarray experiments of gene expression profiling,” *Am. J. Obstet. Gynecol.*, vol. 195, no. 2, pp. 373–388, Aug. 2006, doi: 10.1016/j.ajog.2006.07.001.

[12] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, “Multiple hypothesis testing in microarray experiments,” *Stat. Sci.*, pp. 71–103, 2003.

[13] Y. Hochberg and A. C. Tamhane, *Multiple comparison procedures*. John Wiley & Sons, Inc., 1987.

[14] J. P. Shaffer, “Modified sequentially rejective multiple test procedures,” *J. Am. Stat. Assoc.*, vol. 81, no. 395, pp. 826–831, 1986.

[15] C. E. Bonferroni, “Studi in Onore del Professore Salvatore Ortu Carboni. 1935,” *Calc. delle Assicur. su Grup. di teste*, pp. 13–60.

[16] S. Davis and P. Meltzer, “GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor,” *Bioinformatics*, vol. 23, pp. 1846–1847, Aug. 2007, doi: 10.1093/bioinformatics/btm254.

[17] R. Ihaka and R. Gentleman, “R: A Language for Data Analysis and Graphics,” *J. Comput. Graph. Stat.*, vol. 5, no. 3, pp. 299–314, Sep. 1996, doi: 10.1080/10618600.1996.10474713.

[18] R Core Team, “R: A language and environment for statistical computing and Foundation for Statistical Computing.” 2020.

[19] G. R. Warnes *et al.*, “gplots: Various R programming tools for plotting data,” *R Packag. version*, vol. 2, no. 4, p. 1, 2009.

[20] A. D. Yates *et al.*, “Ensembl 2020,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D682–D688, Nov. 2019, doi: 10.1093/nar/gkz966.

[21] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, “GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists,” *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–7, 2009.

[22] D. W. Huang *et al.*, “DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists,” *Nucleic Acids Res.*, vol. 35, no. suppl\_2, pp. W169–W175, 2007.

[23] B. T. Sherman and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, vol. 4, no. 1, p. 44, 2009.

[24] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[25] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, “REVIGO summarizes and visualizes long lists of gene ontology terms,” *PLoS One*, vol. 6, no. 7, p. e21800, 2011.