Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development

Heba Saadeh, Reem Q. Al Fayez, and Basima Elshqeirat

***Abstract*—Erythropoiesis is the specific lineage in which the haematopoietic stem cells (HSC) differentiate into red blood cells. During their development, HSC undergo global gene expression changes to reflect the current developmental stage needs. A good way to identify the set of genes that have similar global expression patterns across the different developmental stages is through clustering. Unsupervised clustering aims at highlighting these co-regulated genes without prior knowledge regards their full interactions. In this study, we apply k-means clustering on a gene expression microarray data that measures the expression levels of human genes at four erythropoiesis stages. Eight clusters have been identified; one cluster, in particular, of 450 genes (C4) is more active toward the maturation stages and it is involved in cell division and DNA replication processes, which are vital during development. Another cluster of 234 genes (C7) is involved in autophagy (cells consumption/destruction), which is known to be involved in enucleation (expulsion of the nucleus from the cell).**

***Index Terms*—Clustering, elbow method, erythropoiesis, k-means.**

I. Introduction

Haematopoietic stem cells (HSCs) are multipotent stem cells that can develop into various types of blood cells, such as red blood cells, white blood cells, and platelets, realised by conducting different genetic programmes [1]. The specific lineage of developing HSCs into red blood cells (aka erythrocytes) is known as erythropoiesis (from Greek 'erythro' meaning "red" and 'poiesis' meaning "to make" [1]). Human erythrocytes are produced continuously by the proliferation and the differentiation of these multipotent HSCs through the process of erythropoiesis [1].

Erythropoiesis is a specialised process that generates a cell dedicated to the delivery of oxygen to the tissues. As erythroblasts mature, they decrease in size, synthesize more hemoglobin, and undergo membrane reorganization and chromatin condensation [2]. Eventually, the cells expel their organelles before shaping into the biconcave discoid structure [2].

Given successes in developing different types of stem cells to mature tissues *ex vivo*, i.e. in the laboratory outside the body, researchers have tried to develop HSCs into mature red blood cells *ex vivo*. Success in this may potentially lead to laboratory production lines of blood units that replace or reduce dependency on blood donation [3].

Manuscript received September 25, 2019; revised March 22, 2020.

Heba Saadeh and Basima Elshqeirat are with the Department of Computer Science, University of Jordan, Amman, Jordan (e-mail: heba.saadeh@ju.edu.jo, b.shoqurat@ju.edu.jo).

Reem Q. Al Fayez is with the Department of Computer Information Systems, University of Jordan, Amman, Jordan (e-mail: r.alfayez@ju.edu.jo).

Nonetheless, the genetic programmes that drive the maturation of erythrocytes from HSCs are not fully understood. For instance, a phase of rapid growth and replication is triggered in human bodies during intermediate to late stages of erythropoiesis and consequently enables the generation of large amounts of erythrocytes from very small amounts of HSCs [1]. Researchers have not been able to trigger the same genetic programme in the laboratory, which prevents producing commercial amounts of erythrocytes given the limited amounts of HSCs available [3].

Another important step that happens towards the end of erythropoiesis is the enucleation, which is the expulsion of the nucleus from the red blood cell [4], [5]. The nucleus is an important but bulky part of cells. However, it has to be expelled from mature red blood cells to enable them to bend when travelling in the narrowest of the veins. Again, researchers have not been able to realise the enucleation step when running erythropoiesis in the laboratory.

Understanding the pattern of gene expression and identifying the specific genes expressed during each stage of erythropoiesis are vital for a synthesis of erythroid developmental biology. Therefore, in this study, we analyse a gene expression dataset that measures the expression levels of every human gene over four main stages of erythropoiesis, namely, colony-forming unit-Erythroid (CFU-E), pro-erythroblasts (Pro-E), intermediate (Int-E) and Late (Late-E) erythroblasts [6]. The purpose is to apply clustering technique in order to identify interesting sets of genes involved in important biological processes that might enhance the understanding of genes behaviour during these stages.

K-means algorithm is one of the most widely used algorithm in unsupervised learning. Running such explorative experiments on this dataset will emphasis the importance of data mining for exploring gene clusters and their effect on different diseases.

The rest of paper is organised as follows: brief background is presented in the next section covering the main concepts of the paper. The dataset collection and the pre-processing steps needed to prepare the dataset is explained next. After cleaning the dataset, the results of applying K-means on the gene expression microarray data is presented in the following section. Finally, the conclusion and future work are summarised.

II. Background

A. Clustering Gene Expression Patterns

The biological data available for research is increasing exponentially each year [7]. Clustering algorithms is one of

doi: 10.18178/ijmlc.2020.10.3.956

the data mining techniques that is heavily applied on gene expression data. It helps discover the similarity between gene expression profiles and the identification of functionally related genes. Therefore, research in the literature had several studies that focused on surveying clustering algorithms that can be used with gene expression pattern [7]-[12].

From previous studies that emphasised the importance of using clustering algorithms to understand gene expression patterns, it can be deduced that applying clustering techniques on expression datasets identify interesting patterns in the genes that are useful to gain knowledge about the biological processes taking place in an organ [9]. Different clustering algorithms were applied in the literature in this domain. For example, hierarchal clustering of gene expression variations have discovered distinctive gene expression patterns in liver tumour tissues [10]. Another clustering application was performed on diffuse large B-cell lymphomas using two level clustering [11]. The first level is Self-Organizing Maps (SOM) [12], in which the clustering was applied on tumour samples, and then the second level, which is hierarchical clustering and k-means that were used to identify useful patterns of gene expressions.

B. K-means

Clustering a set of genes involved in the different stages of erythropoiesis requires the use of unsupervised learning due to the lack of knowledge regards the relation between the genes and cells development. One of the representative-based clustering algorithms is k-means [13], which aims at partitioning a dataset $\mathbf{D}= \{x_i\}^n{}_{i=1}$ into several **k** clusters denoted by $C= \{C_1, C_2, \ldots, C_k\}$. Each cluster can be summarised by a representative point that is commonly chosen as the centroid (mean: μ_i) of all the points in the cluster. Such algorithms rely on finding the similarity between clusters observations. The k-means method is one of the simplest and most common clustering algorithms [14]. It follows a greedy approach that aims to minimise the sum of squared error (SSE, Eq. 1) over different observations [15].

$$SSE = \sum_{i=1}^{n} (obs_i - exp)^2 \quad (1)$$

C. Cluster Validation

Several experiments can be conducted with k-means to group observations of any dataset into ***K*** number of clusters. Hence, there are relative measures used to compare the validity of such clustering. The goal is to decide on the optimal number of clusters based on comparing such measures. The Elbow method is used for estimating this number. It considers the compactness of the clusters for each clustering experiment by measuring the within sum of squares at different iteration of clustering [16].

III. Dataset Pre-processing

Global gene expression studies of erythroid cells have focused on *in vitro* maturation of immortalised cell lines, primary progenitor cells or *in vivo* derived erythroid cells from a single erythroid lineage [6], [17]-[19]. Many studies assess gene expressions during red blood cells development (Erythropoiesis). One study in particular investigated four developmental stages; CFU-E, Pro-E, Int-E and Late-E [6]. Gene expression for each stage was assessed in three replicates using GeneChip Human Genome U133_plus_2.0 (Affymetrix HGU133_plus_2.0) arrays. The data were deposited in the National Center for Biotechnology Information Gene Expression Onmibus (NCBI GEO) [20].

Under the accession number of GSE22552. The series matrix of the study, which is transformed by log2, was downloaded from GEO and then it was cleaned by removing all the metadata that describes data generation protocols and by deleting all the unnecessary headers. The total number of rows in this dataset is 54,675 (probes: micro holes on the chip) with 13 features (4 stages each with 3 replicates in addition to the probe IDs). The data were pre-processed prior to clustering. The pre-processing steps are shown in Fig. 1.

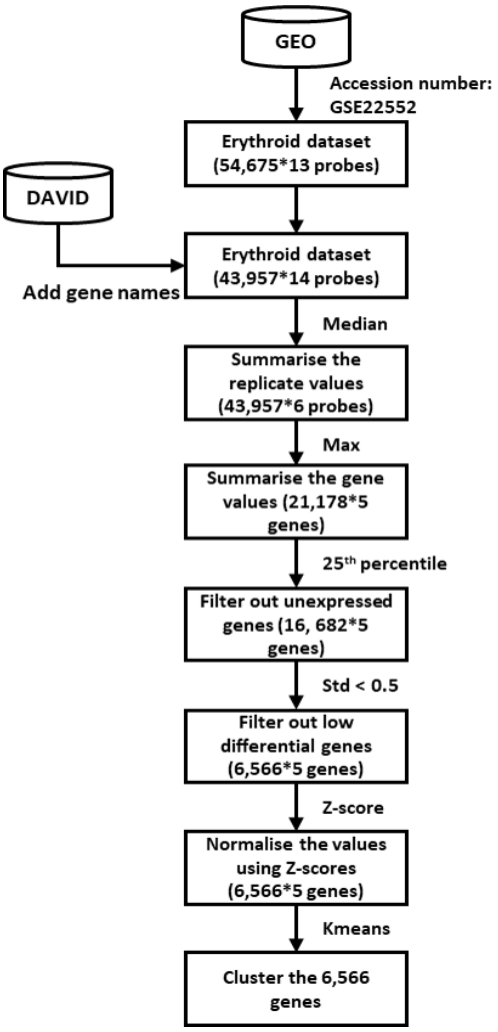
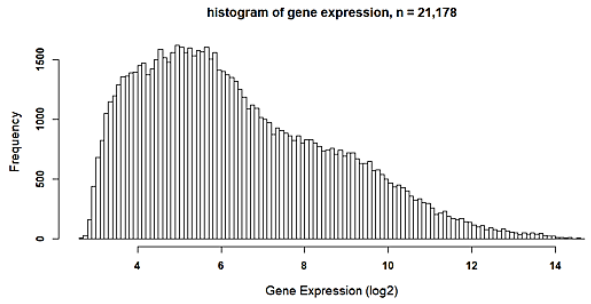


Fig. 1. Dataset pre-processing steps.

In microarrays, each probe ID is mapped to a gene name, and on the array (or chip) a single gene can be represented by more than one probe. There are many tools available for converting from probe IDs to gene names, like DAVID [21]. Some probes represent regions in the genome that do not correspond to a gene, thus these probes were removed. The remaining probes with mapped genes are 43,957.

In machine learning, feature reduction is a vital step specifically if some of the features are redundant. In this dataset, each developmental stage is represented by three values, since the stages were assessed in triplicates, therefore, the three values were summarised into one values, which is

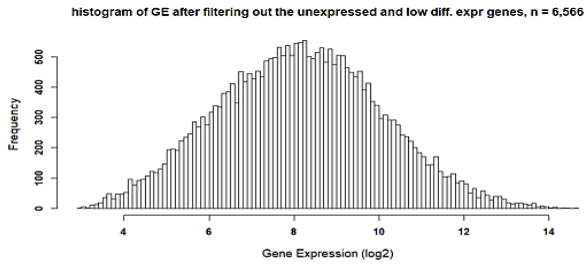
the median, reducing the number of features to 6 (probe ID, gene name and the four stages). Likewise, the different probes that represent the same gene were also summarised and the maximum value among them was selected. The summarised data now contains 21,178 genes and their expression distribution is presented in Fig. 2(a).



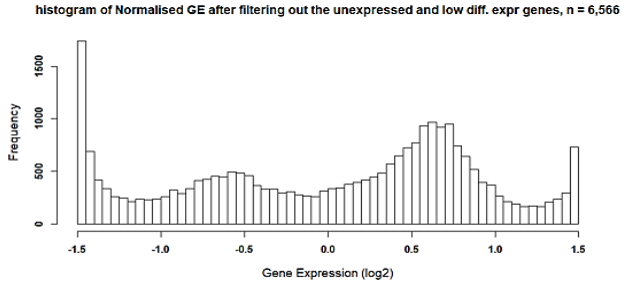
(a)

(b)

Fig. 2. Distribution of gene expressions. (a) Histogram of the data after summarising over rows and columns (n = 2,178). (b) Histogram of the data after filtering the silent genes (n = 16,682).



(a)



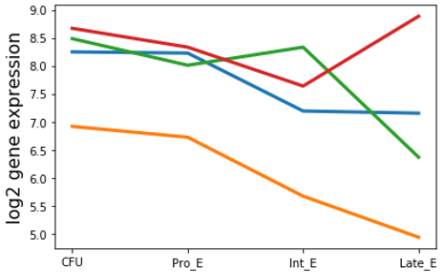
(b)

Fig. 3. Distribution of gene expressions. (a) Histogram of the data after filtering the low differentially expressed genes (n = 6,566). (b) Histogram of the data in (a) after Z-scores normalisation.

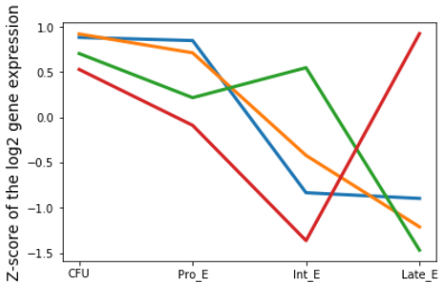
Since the main interest here is on the genes that change their expression across the different stages, the unexpressed genes (genes that show no transcription activity i.e. silent genes) were filtered out. The definition of silent genes was to have low expression levels in all the stages. Low expression was selected to be less than the 25th percentile in all stages. Total number of silent genes is 4,496. The distribution of the remaining genes is shown in Fig. 2(b). The next step was to also exclude the genes that show very little or no differential expression among the different stages. These genes were identified as the ones having small variation; i.e. their standard deviation across the stages is less than 0.5 (Fig. 3(a)).

Removing the low differentially expressed genes shifts the data from being skewed (Fig. 2(b)) to be more normally distributed (Fig. 3(a)). Now in order to expand the variation of gene expression at the different stages the log2 transformed expression levels were normalised using Z-scores (Eq. 2 and Fig. 3(b)). This will have the exact effect needed to highlight the variations, even the smaller ones. Fig. 4 shows the effect of this normalisation for the first four genes.

$$\text{Z-score} = (x - \mu) / \sigma \quad (2)$$



(a)



(b)

Fig. 4. Z-score normalisation effect for the first four genes. (a) Before normalisation. (b) After normalisation. The x-axis shows the different developmental stages and the y-axis shows the expression levels.

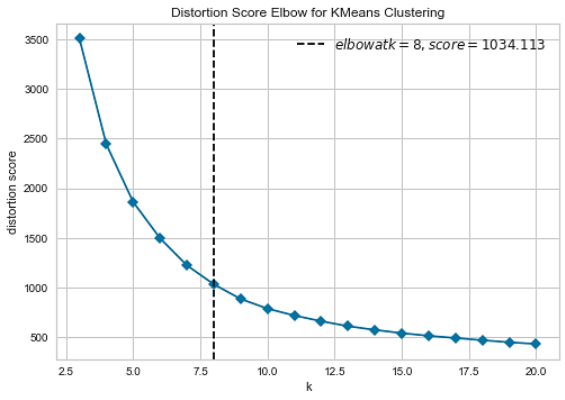


Fig. 5. Distribution of the distortion score for different number of clusters (k) obtained from applying the Elbow method for k-means clustering.

IV. Results and Interpretation

After cleaning, filtering and normalising the data, it

becomes ready for the next step; clustering. K-means is one of the most commonly used clustering techniques, its performance and results are widely acceptable. k-means takes the number of clusters as input, which in some cases is considered a problem, since no prior knowledge might be available. Nevertheless, some methods can help estimating the suitable number of clusters for a certain data. Here, the Elbow method was applied for k-means clustering on the normalised log2 transformed data, and the acceptable number of clusters obtained is 8, as shown in Fig. 5.

After knowing the number of clusters, k-means was applied, and the 8 clusters obtained are shown in Fig. 6. Some clusters showed interesting patterns of groups of genes that change their expression levels at the late erythroid development stage (Late_E), like classes C2, C5 and C8. While other classes highlight the different behaviour of the genes at the intermediate stage (Int_E), like C1 and C7.

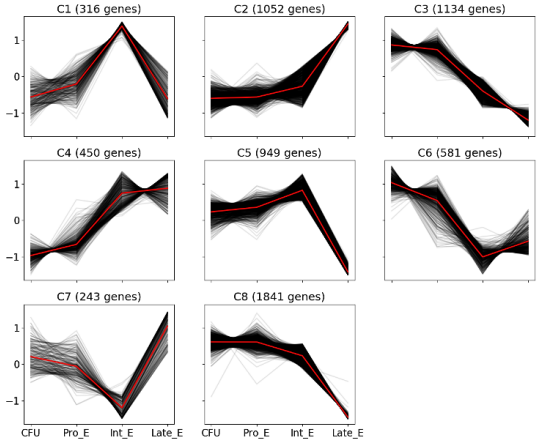


Fig. 6. Line graph of the eight clusters obtained by clustering the data with k-means. The x-axis shows the different developmental stages and the y-axis shows the z-scores "normalised log2 gene expression". The number of cluster along with the total number of genes in each cluster is presented.

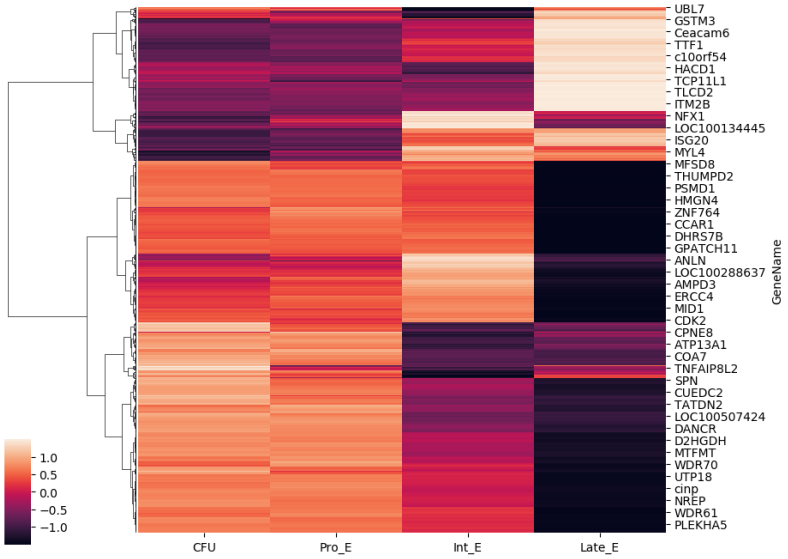


Fig. 7. Heatmap for gene expression (in Z-scores) across the different developmental stages. A dendrogram of some of the genes is shown on the left side of the map. Selected gene names are shown on the right. The key to the colours used is shown in the top left.

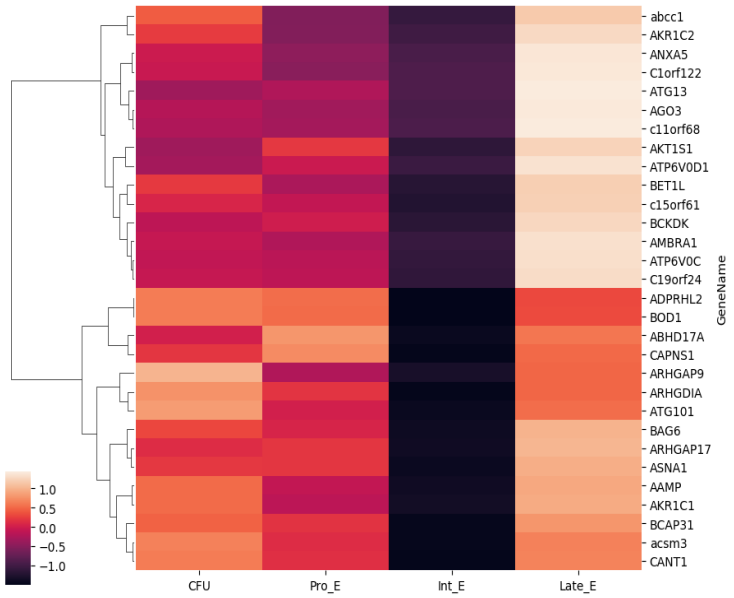


Fig. 8. Heatmap for gene expression (in Z-scores) across the different developmental stages for the first 30 genes in cluster 1.

Another way to visualise gene expression across the different stages is through heatmaps. Fig. 7 shows a heatmap for a random set of genes, which highlights gene patterns from different clusters, while Fig. 8 demonstrates the first 30 genes from cluster 7, in which the genes express lower expression levels at the Int-E stage compared to other stages.

Some of the genes in this cluster are enriched with autophagy (self-consumption); which is a process by which

cytoplasmic components and organelles are degraded [22], [23]. Autophagy is known to be involved in enucleation (expulsion of the nucleus from the cell).

The genes in cluster 3 are more enriched in RNA processing and translation. While cluster 4 shows the genes that increase their transcription during cells growth, thus, the biological processes that this cluster are involved in are growth-related, like DNA replication and cell division. Summary of gene ontology terms, obtained from DAVID [21] is presented in Table I.

TABLE I: GENE ONTOLOGY BIOLOGICAL PROCESSES TERMS

Cluster	GO ID	GO Term	P-value	Bonferroni	FDR
C1	GO:0045926	Negative regulation of growth	7.3E-06	8.9E-03	0.01
	GO:0071294	Cellular response to zinc ion	7.3E-06	8.9E-03	0.01
C3	GO:0070125	Mitochondrial translational elongation	8.4E-09	2.5E-05	1.5E-05
	GO:0006364	rRNA processing	5.6E-08	1.6E-04	1.0E-04
	GO:0070126	Mitochondrial translational termination	5.7E-08	1.7E-04	1.0E-04
	GO:0008033	tRNA processing	5.6E-07	1.6E-03	1.0E-03
	GO:0008380	RNA splicing	2.5E-05	7.1E-02	0.04
C4	GO:0051301	Cell division	1.4E-28	3.3E-25	2.4E-25
	GO:0007067	Mitotic nuclear division	5.6E-22	1.3E-18	9.8E-19
	GO:0006260	DNA replication	8.2E-19	2.0E-15	1.4E-15
	GO:0007062	Sister chromatid cohesion	6.0E-17	2.7E-13	2.0E-13
	GO:0006281	DNA repair	7.0E-15	1.7E-11	1.2E-11
C6	GO:0030168	Platelet activation	1.2E-05	2.5E-02	0.02
C7	GO:0006914	Autophagy	1.1E-05	1.2E-02	0.02

V. CONCLUDING REMARKS

Unsupervised clustering is widely applicable to various domains in order to uncover interesting and hidden patterns and similarities. In this study, k-means clustering was applied on human erythropoiesis, which is a specific lineage of developing haematopoietic stem cells into red blood cells, gene expression dataset that measures global gene activity across 4 different developmental stages; CFU-E, Pro-E, Int-E and Late-E. Using the Elbow method, the optimal number of clusters was estimated to 8.

The identified clusters illustrate distinct interesting patterns, which might contribute to the understanding of the genetic programmes that drive the maturation of erythrocytes from HSCs. Cluster 4, in particular, is related to a phase of rapid growth and replication that is triggered in human bodies during intermediate to late stages of erythropoiesis and consequently enables the generation of large amounts of erythrocytes from very small amounts of HSCs. While cluster 7 is enriched with a process called autophagy. This process is known to be involved in enucleation (the expulsion of the nucleus from the red blood cell), which gives the mature red blood cells the ability to bend when travelling in the narrow veins.

Understanding the general behaviour of genes during each stage of erythroblasts development might assess the *ex vivo* studies that aim to produce them in the labs, which can replace or reduce dependency on blood donation.

Besides K-means, other clustering techniques like, K-median and hierarchical clustering can also be applied to similar datasets. Furthermore, applying clustering with deep learning may uncover hidden patterns that traditional techniques did not uncover.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

H. S. conducted the research, analysed the data and wrote the paper, R. Q. A., wrote and revised the paper, and B. E. has revised the paper. All authors had approved the final version.

REFERENCES

[1] M. Moras, S. D. Lefevre, and M. A. Ostuni, "From erythroblasts to mature red blood cells: Organelle clearance in mammals," *Front Physiol*, vol. 8, p. 1076, 2017.

[2] J. J. Welch, J. A. Watts, C. R. Vakoc *et al*., "Global regulation of erythroid gene expression by transcription factor GATA-1," *Blood*, vol. 8104, no. 10, pp. 3136–3147, 2004.

[3] A. P. Ng and W. S. Alexander, "Haematopoietic stem cells: Past, present and future," *Cell Death Discov*., vol. 83, p. 17002, 2017.

[4] L. Blanc, A. D. Gassart, C. G éminard, P. Bette-Bobillo, and M. Vidal, "Exosome release by reticulocytes — an integral part of the red blood cell differentiation system," Blood Cells. Mol. Dis. 2005, vol. 835, pp. 21–26.

[5] G. Keerthivasan, A. Wickrema, and J. D. Crispino, "Erythroblast enucleation," *Stem Cells Int*, p. 139851, 2011.

[6] A. T. Merryweather-Clarke, A. Atzberger *et al*., "Global gene expression analysis of human erythroid progenitors," *Blood*, vol. 8117, no. 13, pp. e96-e108, 2011.

[7] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. J. Brown "Incremental genetic K-means algorithm and its application in gene expression data analysis," *BMC Bioinformatics*, vol. 85, no. 1, p. 172, 2004.

[8] J. H. Do and D. Choi, "Clustering approaches to identifying gene expression patterns from DNA microarray data," *Molecules and Cells*, vol. 825, no. 2, p. 279, 2008.

[9] J. Oyelade, I. Isewon, F. Oladipupo *et al*., "Clustering algorithms: Their application to gene expression data," *Bioinformatics and Biology Insights*, 2016.

[10] X. Chen, S. T. Cheung *et al*., "Gene expression patterns in human liver cancers," *Molecular Biology of the Cell*, vol. 813, no. 6, pp. 1929-1939, 2002.

[11] J. Wang, J. Delabie, H. C. Aasheim, E. Smeland, O. and Myklebost, "Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study," *BMC Bioinformatics*, vol. 83, no. 1, p. 36, 2002.

[12] T. Kohonen, *Self-organizing Maps*, Berlin, Springer, 1997, vol. 8117.

[13] C. C. Aggarwal, *An Introduction to Data Mining*, Cham: Springer International Publishing, 2015.

[14] M. A. Syakur, B. K. Khotima, E. M. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for Identification of the best customer profile cluster," in *Proc. IOP Conf. Series: Materials Science and Engineering*, 2018

[15] M. J. Zaki and W. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*, 2014.
[16] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud*., vol. 1, no. 6, 2013.
[17] A. N. Gubin, J. M. Njoroge, G. G. Bouffard, and J. L. Miller, "Gene expression in proliferating human erythroid cells," *Genomics*, vol. 859, no. 2, pp. 168–177, 1999.
[18] P. Wong, S. M. Hattangadi, A. W. Cheng, G. M. Frampton, R. A. Young, and H. F. Lodish, "Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes," *Blood*, vol. 8118, no. 16, pp. e128–e138, 2011.
[19] J. Isern, Z. He *et al*., "Single-lineage transcriptome analysis reveals key regulatory pathways in primitive erythroid progenitors in the mouse embryo," *Blood*, vol. 8117, no. 18, pp. 4924–4934, 2011.
[20] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res*., vol. 830, no. 1, pp. 207-210, 2002.
[21] W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat Protoc*., vol. 84, no. 1, pp. 44-57, 2009.
[22] N. Mizushima, "Autophagy: Process and function," *Genes Dev.*, vol. 821, pp. 2861–2873, 2007.
[23] D. J. Klionsky, "Autophagy revisited: a conversation with Christian de duve," *Autophagy*, vol. 84, no. 6, pp. 740–743, 2008.

Heba Saadeh received the B.Sc. in 2007 and M.Sc. in 2009 in computer science from the University of Jordan, and the Ph.D. degree in 2014 in bioinformatics from King's College London, UK. Her research interests include epigenetics and bioinformatics and also has interests in machine learning in medical applications. Currently she is working as an assistant professor at The University of Jordan.

Reem Q. Al Fayez received her B.Sc. degree in 2010 in computer information systems from the University of Jordan, and her M.Sc. degree in 2012 in information systems from the University of Jordan. In 2016, she received her Ph.D. degree in computer science from the University of Warwick, UK majoring in Ontologies and Linked Data for databases integration. Her research interests include graph databases and its application in data management. Currently she is working as an assistant professor at The University of Jordan.

Basima Elshqeirat received the B.Sc. in 2004, and M.Sc. in 2008 degrees in computer science from the University of Jordan, and the Ph.D. degree in 2015 in computer science from Curtin University, Australia. Her research interests include network reliability and network design. Also she has research interests in meta-heuristic algorithms, genetic algorithms and optimization problem. Currently working as an assistant professor at The University of Jordan.