SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Identifying Transcriptional Pattern through Clustering Analysis of Gene Expression Data

Shamini Raja Kumaran[1,*], Ding Daorui[2], Chen Yanhao[2], Hong Chang[2], Bi Xiaoyang[1], Valarmathie Gopalan[2], Shaidah Jusoh[1], Arda Yunianta[3]

[1] School of Electrical Engineering and Artificial Intelligence, Xiamen University Malaysia,Jalan Sunsuria, Bandar Sunsuria, 43900, Sepang, Selangor
[2] School of Computing and Data Science, Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900, Sepang, Selangor
[3] King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

The outbreak of SARS-CoV-1 in 2002 followed by SARS-CoV-2 in 2019 resulted in a global health crisis, emphasizing the need to understand the molecular basis of the viral infection. This study aimed to investigate transcriptional pattern in SARS-CoV-1 through clustering analysis, using the Gene Expression Omnibus microarray data set. By applying integrated hierarchical clustering and k-medoid methods, we sought to elucidate the transcriptional patterns associated with SARS-CoV-1 infection and draw significant conclusions regarding viral pathogenesis. Our analysis revealed distinct clusters of genes with similar expression patterns, providing insights into the host's response to SARS-CoV-1 infection. Notably, key genes involved in viral replication, immune response, and host-pathogen interactions exhibited significant alterations in expression levels. Additionally, the clustering analysis unveiled subgroups within the infected samples, implying potential variations in the host response or viral strain differences. In light of these findings, we conclude that gene expression profiling with clustering analysis, offers valuable insights into the molecular dynamics of the SARS-CoV-1 outbreak. The identified transcriptional patterns enhance our understanding of the virus-host interaction and may pave the way for the identification of potential therapeutic targets. Ultimately, this research contributes to a comprehensive understanding of the pathogenesis of SARS-CoV-1 and provides a basis for future investigations into effective intervention strategies such as gene ranking and gene selection toward SARS-CoV-2.

Keywords:
Gene expression; clustering analysis; SARS-CoV

1. Introduction

The severe acute respiratory syndrome coronavirus (SARS-CoV-1) and SARS-CoV-2 epidemics that occurred in recent years have highlighted the necessity to understand the molecular mechanisms behind viral illnesses. Significant changes in host gene expression patterns can result from viral infections, indicating the complex interplay between the virus and the host cellular

* *Corresponding author.*
E-mail address: shamini.rajakumaran@xmu.edu.my

https://doi.org/10.37934/araset.59.1.111

response hybridized with clustering analysis. This approach allows us to decipher the transcriptional changes associated with viral infections and uncover crucial insights into viral pathogenesis.

The previous research studies demonstrate the degree of similarity between various diseases, is crucial for understanding and appreciating disease biology that allowed us to conduct this research. The research [1] focused on identifying distinct transcriptional patterns associated with SARS-CoV-1 infection using clustering analysis. Researchers identified a several clusters with various expression patterns by evaluating gene expression data from patient samples using hierarchical clustering. Functional analysis revealed enrichment of genes involved in immune response, cytokines signaling, and viral replication. The findings suggested heterogeneity in host responses and provided insights into potential therapeutic targets. Thair *et al.,* [2] investigated the dynamic transcriptional changes during SARS-CoV-1 infection in macrophages. The research utilized RNA sequencing to identify differentially expressed genes at different time points after infection. Clustering analysis uncovered distinct expression profiles associated with various immune response pathways, such as inflammation and antiviral defense. It is revealed important new information on the host-virus interaction in macrophages and demonstrated the dynamic nature of gene expression during SARS-CoV-1 infection [2]. Even more recent investigations, such as Ramesh *et al.,* [3], made use of the same data set to determine the host immune response to the pathogenic corona-virus, which in turn lowers mortality. In the near future, an immunological intervention for COVID-19 might be developed that would lower the death rate, according to this study article's novel perspectives [3].

By examining the changes in gene expression, transcriptional patterns, and host responses over the course of infection, this research advances our understanding of the SARS-CoV-1 infection. These studies investigate the SARS-CoV-1 transcriptional patterns using cluster analysis and offer insightful information. One of the most popular unsupervised learning techniques is clustering, which divides data into groups depending on how similar the examples are to one another [4]. While clustering method is a valuable tool for grouping genes based on their expression patterns, enabling the identification of co-regulated genes and uncovering biological insights [5]. In this study, we employed hybrid hierarchical clustering and partitional clustering methods to identify distinct clusters of genes with similar expression profiles. These clustering approaches can reveal transcriptional patterns associated with viral infection, shedding light on key biological processes and pathways involved in SARS-CoV-1 pathogenesis [6]. By characterizing the transcriptional patterns associated with viral infection, we aim to unravel the molecular mechanisms driving SARS-CoV-1 pathogenesis and gain insights that have implications for combating other corona viruses.

The aim of the research paper is to investigate gene expression in SARS-CoV-1-infected samples using the Gene Expression Omnibus dataset GSE1739, which comprises microarray data specifically related to this viral infection. By analyzing the expression levels of thousands of genes simultaneously, we can identify transcriptional patterns that are indicative of the host response to SARS-CoV-1 infection. Hereby, hierarchical clustering known as agglomerative clustering is a type of clustering algorithm that follows a bottom-up approach. It starts by considering each data point as an individual cluster and iteratively merges the most similar clusters until all data points belong to a single cluster or until a specified number of clusters (k) is reached [7].

Partitional clustering is a type of clustering algorithm that aims to partition the dataset into a predetermined number of non-overlapping clusters [8]. Unlike hierarchical clustering, partitional clustering directly assigns each data point to a specific cluster based on certain criteria. Therefore, in this research, hybridizing hierarchical and partitional clustering, also known as hybrid clustering, is a combination of both hierarchical and partitional clustering methods. This approach seeks to leverage the strengths of both techniques to overcome their respective limitations and improve the interpretability of gene profiling results. Table 1 shows the difference between the categories of

clustering techniques and provides information on convergence rate, scalability, time complexity, and computing efficiency. Therefore, the idea behind gene clustering for this research is to utilize the power of both partition and agglomerative clustering methods.

Table 1
Different clustering techniques

Categories	Scalability	Time complexity	Convergence rate	Efficiency
Hierarchical	High	High	Low	High
Partitional	Low	Low	Low	High

The remainder of this article is organized as follows: Section 2 describes the materials and methods, including data preprocessing, and the implementation of hierarchical agglomerative clustering (HAC) and k-Medoids clustering. In Section 3, we present the results of our clustering analysis and discuss the biological significance of the identified gene expression patterns. Finally, Section 4 summarizes the study with a summary of this research and future research. Beyond SARS-CoV-1, the conclusions of this investigation have a broader impact. The research will provide valuable insights into the molecular mechanisms of similar corona viruses and infectious diseases, such as SARS-CoV-2. The understanding acquired by researching SARS-CoV-1 gene expression profiling can assist in improving diagnostic procedures, and help us be more prepared for viral epidemics in the future.

2. Methodology

2.1 Understanding Microarray Dataset

The microarray gene expression data set, GSE1739, were retrieved from the Gene Expression Omnibus (GEO) database of NCBI (https://www.ncbi.nlm.nih.gov/geo/). Table 2 indicates the information of GSE1739. GSE1739 is a gene expression dataset identifier in the Gene Expression Omnibus (GEO) database. The data set was generated using the GPL201 Affymetrix Human HGFocus Target Array platform. It consists of gene expression data obtained from 14 samples, including 10 peripheral blood samples from adult patients with SARS and 4 control samples. The data set aims to explore gene expression differences between patients with SARS and healthy controls, likely to gain insights into the molecular mechanisms and gene expression changes associated with SARS infection.

Table 2
GSE1739's information

No. of controls	4
No. of patients	10

The gene expression analysis between the test and control samples were analyzed using an online analysis tool which is GEO2R for the mentioned data, GSE1739. Figure 1 shows that each point on the plot represents an individual data point of gene. Based on Figure 1, the clusters of points that appear close together on the UMAP plot represent groups of data points with similar features or attributes. Each cluster might correspond to a distinct subgroup or category within the data set.

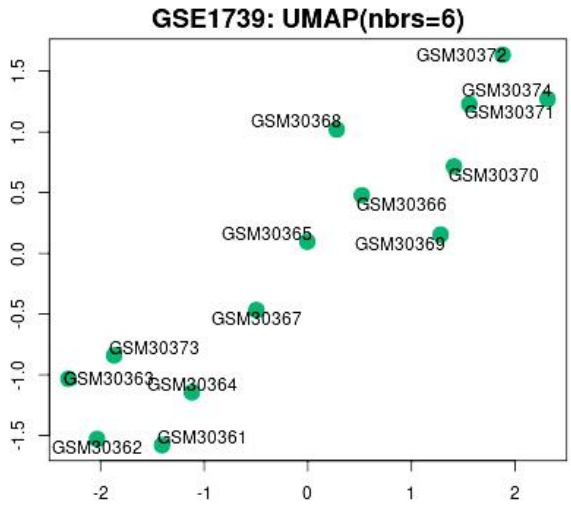

Fig. 1. UMAP plot of GSE1739

Figure 2 shows the differentially expressed genes value distribution of datasets GSE1739. The analysis of GSE1739 of the samples respectively with a set cut-off criterion |logFC|≥1.0 and p < 0 .05.

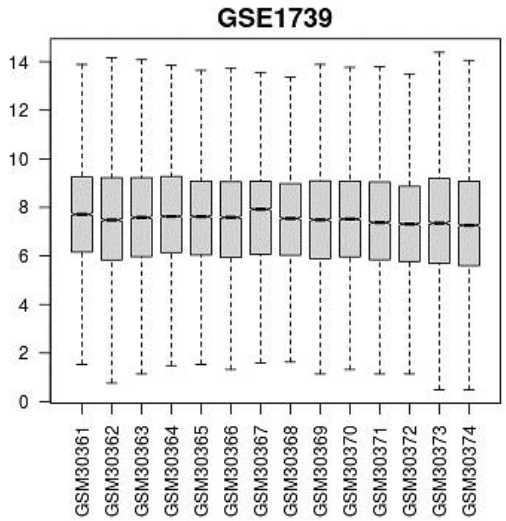

Fig. 2. Boxplot of GSE1739

2.2 Research Process

Figure 3 shows the general methodology of this research article. As it can be seen, the data collected and pre-processed initially, then, gene manipulation approach will be implemented on the collected GSE1739 in order to identify transcriptional patterns that are indicative of the host response to SARS-CoV-1 infection.

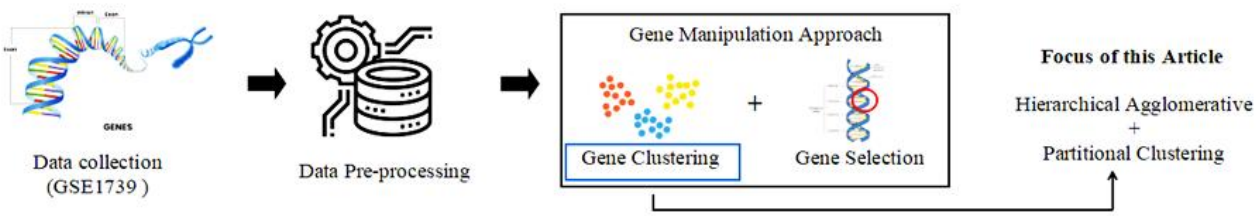

Fig. 3. Methodology of the research

The partitional clustering that will be utilized in this research is k-medoid. Figure 4 shows the pseudo code for the developed integrated hierarchical and k-medoid clustering algorithm. The pseudo code integrates the HAC with k-medoid clustering to effectively group genes based on their expression patterns. It iteratively merges clusters in a way that optimizes the overall similarity within clusters and dissimilarity between clusters. The final result is the best clustering configuration that reveals underlying patterns in the gene expression data, helping to better understand gene interactions and potential biological implications. In addition, using these clustering methods, you can identify the groups of genes that exhibit similar expression patterns, functional characteristics, or other shared features. It is important to note that the effectiveness of clustering depends on the choice of distance metrics, linkage criteria (for hierarchical clustering), and the number of clusters (*k*) in k-medoid clustering, whereby, for this research, the choice were Euclidean distance, average linkage, and two clusters.

```
Initialize the nodes list with individual data points as clusters.
Calculate the distance between all pairs of nodes and store them in a list.
While the number of clusters is more than the desired outcome clusters:
    Find the two closest nodes from the distance list.
    Merge the two closest nodes into a new cluster.
    Update the distance list with the distances to the new cluster.
Perform k-medoid clustering on the obtained clusters.
Calculate the silhouette score for each k-medoid clustering configuration.
Save the best clustering configuration with its medoids and labels.
```

Fig. 4. Pseudo code of integrated hierarchical and k-medoid clustering

To discuss further, there are a certain drawback to depend only on a single method, such as HAC or k-medoid clustering. While HAC may create a hierarchical gene cluster structure that shows different similarity levels, it may be difficult to identify the key representatives within each cluster. Similar to HAC, utilizing k-medoid clustering alone can reveal unique gene profiles for each cluster but may fall short in capturing the full hierarchical linkages across clusters. The hybrid clustering strategy created in this study, however, turns out to be the best option. It addresses the drawbacks of each individual technique by combining the advantages of both HAC and k-medoid clustering. This hybrid approach explicitly identifies important representatives within clusters in addition to revealing hierarchical gene relationships. Consequently, the hybrid clustering technique emerges as the preferable and most useful choice when attempting to get a comprehensive knowledge of gene expression patterns. The importance of employing a HAC, k-medoid, or hybrid technique in a given situation relies on the objectives of the study and the nature of the data.

Table 3 shows a brief overview of the three clustering methods, emphasizing the mathematical perspective, implementation, aim and drawbacks of each clustering methods. It shows that selecting the most appropriate approach depends on a number of variables, including the data's structure, the amount of information sought in clustering, and the findings. Each approach has advantages and disadvantages. Therefore, a hybrid method is crucial for our research since the genes offer complicated information about SARS-CoV-1 infection.

Table 3
Comparison between singular and hybrid clustering methods

Characteristics	HAC	k-medoid	Hybrid
Mathematical perspective	Average linkage between these pseudo-nodes and other gene measurements is calculated. This process is repeated until only one node remains. The distance is using Euclidean distance.	Calculate the distance between the pair of genes using the similarity measures using Euclidean distance. Assign the nearest medoids and obtain the cluster results.	Refer to Figure 4 for the mathematical perspective.
Implementation	Create initial clusters, *n*, each containing a data point, then iteratively combine the closest clusters.	Partition the data into *k* clusters, then utilize k-medoid to get the medoid for each cluster.	Utilize the k-medoid to discover the medoid for each cluster after applying HAC to *k* levels to produce *k* clusters.
Aim	Hierarchical relationships between genes.	Central gene's representations.	Hybridization between hierarchical structure and medoid representation for clustering.
Limitations	Unable to identify the central points within each cluster.	Ignores the hierarchical relationships between genes	Complex approach due to its hybrid nature between hierarchical and partitional clustering methods

3. Results

Analyzing the transcriptional patterns using the developed method provides valuable insights into the interactions between SARS-CoV-1 and the host cells, as demonstrated with the GSE1739 data resulting in two distinct clusters. The first cluster, obtained through hierarchical and k-medoid clustering, reveals a diverse set of genes involved in various cellular functions. Notably, it includes ribosomal proteins essential for protein synthesis, indicating potential increased protein production during viral infection. Moreover, the presence of genes related to RNA processing and splicing suggests active regulation of gene expression in response to the viral challenge. The cluster also contains genes associated with the immune response and anti-oxidative stress response, indicating the host's efforts to counteract viral infection and minimize cellular damage. Additionally, the genes involved in cellular homeostasis, vesicular trafficking, and endosomal functions may play roles in viral entry and replication processes. By studying the transcriptional patterns of these genes in the context of SARS-CoV-1 infection, we can gain insights into how the virus may hijack or alter these cellular processes to promote its replication and evade the host's immune response. In order to understand, the transcriptional patterns of the genes. The genes have been clusteres into two clusters using the developed integrated hierarchical and k-medoid clustering algorithm. Figure 5 shows the excerpts of results of first and second cluster.

ID	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303	GSM303	GSM303	GSM303	GSM303	Gene Syn
1007_s_at	0.0142568	0.0139203	0.0166151	0.0249128	0.0045155	0.0022565	0.0017621	0.004849	0.0109543	0.0051148	0.0105657	0.0044374	0.0122185	0.0077647	DDR1 MIR4
1053_at	0.009062	0.0158973	0.0109041	0.0177696	0.0047507	0.0026783	0.0003611	0.0031383	0.0076912	0.0029517	0.010162	0.0035621	0.0027993	0.0089805	RFC2
117_at	0.0238827	0.0198893	0.026569	0.024593	0.0124127	0.0036877	0.0018539	0.0161054	0.0242989	0.0128089	0.0620647	0.0315985	0.0107306	0.024827	HSPA6
121_at	0.056703	0.0477626	0.051846	0.0704455	0.0175118	0.0102691	0.0079073	0.0184661	0.0446641	0.021603	0.0459121	0.024533	0.0298087	0.0438746	PAX8
1255_g_at	0.0022466	0.0023463	0.0020612	0.0017552	0.0011494	0.0003611	0.0001313	0.0011197	0.0021625	0.0009106	0.0034125	0.0007354	0.0014459	0.0043346	GUCA1A
1294_at	0.0286113	0.0322129	0.0338535	0.047737	0.0139016	0.0051614	0.0018899	0.0107058	0.0391912	0.0205901	0.0296464	0.0085589	0.0359285	0.0266654	MIR5193 U
1316_at	0.0084449	0.0093255	0.0064299	0.0135313	0.0063306	0.002038	0.0029407	0.0046189	0.00457	0.00403	0.0102966	0.0053285	0.0071538	0.0095972	THRA
1320_at	0.0003863	0.000277	0.0004575	0.0013946	0.0003262	0.000176	0.0004577	0.0003017	0.000503	0.0004313	0.000845	0.0002842	0.0003447	0.0006461	PTPN21
1431_at	0.0039072	0.0027537	0.0048211	0.0063608	0.001926	0.0001477	0.0005468	0.0011835	0.0021711	0.001185	0.0028527	0.0004016	0.0023075	0.0021086	CYP2E1
1438_at	0.0107359	0.0010591	0.0019254	0.0018912	0.0028203	0.0003924	0.0020015	0.001311	0.0039552	0.0015859	0.0055923	0.0003203	0.0020932	0.0009809	EPHB3
1487_at	0.0255255	0.0155877	0.0200336	0.0331172	0.0079637	0.0034328	0.0009127	0.0048724	0.0237228	0.008487	0.0209053	0.0111329	0.0122732	0.0185424	ESRRA
1494_f_at	0.0099012	0.0069846	0.0113666	0.0144497	0.0034726	0.0002023	0.0001911	0.0049548	0.0078417	0.0032327	0.0084235	0.0056962	0.0060988	0.0139377	CYP2A6
1598_g_at	0.0137462	0.0181024	0.0156046	0.0260897	0.00448	0.0030647	0.0005515	0.0014479	0.0186498	0.0132271	0.0184563	0.0098425	0.0097009	0.0097147	GAS6
160020_at	0.0354444	0.0200577	0.0158861	0.0178648	0.0111324	0.0031709	0.0046735	0.0053762	0.014342	0.0074261	0.0139512	0.011724	0.0107643	0.0223132	MMP14
1729_at	0.0434186	0.0389259	0.0304651	0.0438184	0.0137019	0.0072773	0.0023272	0.0120153	0.0374372	0.0126172	0.018564	0.0107088	0.0337765	0.0176203	TRADD
1773_at	0.0084759	0.0086574	0.0061735	0.0039594	0.0038654	0.0006605	9.93E-05	0.0025551	0.0128416	0.0020804	0.0031272	0.0025853	0.0035685	0.0074651	CHURC1-FN
177_at	0.0062471	0.0067837	0.0049317	0.0088848	0.003162	0.0011379	0.001401	0.0037511	0.0062166	0.0028885	0.0075139	0.0042276	0.0044848	0.0086633	PLD1
1861_at	0.0047996	0.0087769	0.0088982	0.0116876	0.004409	0.0015445	6.46E-05	0.0025349	0.0092948	0.004176	0.008585	0.0044825	0.007738	0.0121287	BAD
200021_at	0.5005772	0.6160907	0.4415231	0.5486996	0.4367872	0.3033048	0.4567744	0.3041044	0.5564632	0.4170472	0.6009904	0.4599598	0.5017884	0.6519106	RPL18
200035_at	0.031111	0.0357541	0.0310935	0.0374508	0.0211419	0.0130809	0.0061126	0.0184459	0.0573079	0.0285804	0.0481673	0.0201205	0.0477309	0.0599619	RPL10A

(a)

ID	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303€	GSM303	GSM303	GSM303	GSM303	GSM303	Gene Syn
200000_s_a	0.0677453	0.0592552	0.0772185	0.1084678	0.025693	0.0083444	0.0015091	0.018858	0.0577895	0.0245765	0.0289036	0.0139145	0.0612272	0.0548286	PRPF8
200001_at	0.0918322	0.1168049	0.1738724	0.2459233	0.0642868	0.0378004	9.25E-05	0.0434468	0.2564025	0.1034427	0.1704236	0.056497	0.1228285	0.1696132	CAPNS1
200002_at	0.3607399	0.451105	0.4853305	0.4722946	0.2125711	0.0893635	0.0102645	0.1550366	0.3135514	0.1758565	0.2495775	0.069579	0.3891105	0.2792644	RPL35
200003_s_a	0.5729623	0.7195998	0.7347624	0.7818331	0.443921	0.220199	0.0238328	0.2653698	0.5698551	0.4087693	0.4588945	0.2166193	0.6160804	0.5090745	MIR6805 R
200004_at	0.2518271	0.2684405	0.2250699	0.2903608	0.1092179	0.0455329	0.041946	0.0500267	0.1663564	0.0464344	0.2359438	0.0622203	0.239075	0.2256693	EIF4G2
200005_at	0.1189694	0.131176	0.1739126	0.2295075	0.0625405	0.0225298	0.0002897	0.034234	0.111456	0.0787224	0.0867646	0.0331235	0.1527171	0.1172573	EIF3D
200006_at	0.1392646	0.2350273	0.2634881	0.3323832	0.1487015	0.0573981	0.0262872	0.0933818	0.2256506	0.1428911	0.3224608	0.0837552	0.2640921	0.2537032	PARK7
200007_at	0.21573	0.197318	0.233204	0.2631622	0.1165159	0.0421658	0.0282417	0.051669	0.1955005	0.1085924	0.1869638	0.0884475	0.1753889	0.220994	SRP14
200008_s_a	0.0930887	0.1533845	0.143543	0.1863218	0.0284311	0.0176101	0.0044851	0.0230647	0.0742596	0.0148762	0.0629259	0.0234931	0.055347	0.081958	GDI2
200009_at	0.1841528	0.2546234	0.2444499	0.3221038	0.1298805	0.0583387	0.0054128	0.0861627	0.1697657	0.0651293	0.1762797	0.0903831	0.208972	0.2241481	GDI2
200010_at	0.4598492	0.5444196	0.5460999	0.5180111	0.3231536	0.1241155	0.0121347	0.1813345	0.3212641	0.2715834	0.352737	0.1051412	0.4250516	0.3840701	RPL11
200011_s_a	0.052747	0.0575171	0.0638812	0.0854871	0.0309984	0.0104481	0.0018614	0.0125316	0.0612632	0.0222761	0.0787287	0.0366067	0.0491306	0.1124059	ARF3
200012_x_a	0.528891	0.5502311	0.5617547	0.557217	0.4203693	0.2326498	0.1152139	0.2883505	0.3343078	0.237727	0.4364067	0.232167	0.5200343	0.4424403	RPL21 RPL2
200013_at	0.4391944	0.5380488	0.6200356	0.6557455	0.2568299	0.1135622	0.0048299	0.1390727	0.2302593	0.2008557	0.3701975	0.1148732	0.495656	0.3042383	RPL24
200014_s_a	0.052858	0.1022871	0.1017666	0.1462042	0.0327492	0.0119795	0.0196509	0.0275296	0.0480862	0.0256679	0.104645	0.0255053	0.0887663	0.1009879	HNRNPC
200016_x_a	0.5124053	0.5297444	0.5350801	0.577769	0.2876707	0.1251795	0.2475713	0.155528	0.2827994	0.2100811	0.3633242	0.123008	0.4530403	0.4967167	RPS27A
200017_at	0.5746406	0.4359246	0.6712532	0.7490357	0.3020272	0.0978424	0.0971618	0.1642245	0.3720158	0.1839906	0.4340923	0.1409109	0.5074291	0.452061	LOC100508
200018_at	0.7176525	0.5647435	0.5857045	0.5698231	0.3626816	0.2081761	0.4942618	0.3261941	0.4334897	0.3477092	0.4587384	0.2256452	0.4904568	0.4224119	FAU
200019_s_a	0.3509542	0.568176	0.5587886	0.7032308	0.3711047	0.1655976	0.0495507	0.2591257	0.5527401	0.3523775	0.4738038	0.2105441	0.5143096	0.4931633	TARDBP
200020_at	0.0716658	0.0849722	0.0735386	0.1068418	0.0232699	0.0101427	0.0059228	0.0116576	0.0353263	0.0174031	0.0540826	0.0224441	0.0677169	0.0771594	CFL1
200022_at	0.5995223	0.5056186	0.5647811	0.7618662	0.3525877	0.1238819	0.0090464	0.1368115	0.5197783	0.2871393	0.3585984	0.1374504	0.493399	0.425701	EIF3F

(b)

Fig. 5. (a) Excerpt result of first cluster (b) Excerpt result of second cluster

First cluster demonstrates strong effects of the virus on host cellular processes and is anticipated to hold important information regarding SARS-CoV-1 and its consequences. For instance, the virus interacts with genes including RPL14, RPL19, and RPS11 to regulate the host's ribosomes for viral protein production [9]. The host may suffer serious repercussions as a result of this disturbance of regular cellular processes including protein synthesis. Additionally, it has been discovered that SARS-CoV-1 alters gene expression by focusing on genes like YY1 and HNRNPC, promoting its replication and thwarting the host's immune response, as demonstrated with GDI2 and PRPF8 [10]. In addition, the virus causes endoplasmic reticulum stress as a result of viral protein buildup, a process that involves genes including XBP1 and HSPA1A. Moreover, SARS-CoV-1 manipulates cellular processes like cell adhesion, facilitated by genes like ACTN4 and CLTC, to aid in viral entry and replication [11]. The disruption of cell cycle and DNA repair mechanisms caused by SARS-CoV-1, as evidenced by the effects on genes such as TSG101 and CDC45, leads to genomic instability that may contribute to the development of diseases, including cancer [12-14]. Understanding these disruptions is critical in devising effective strategies to combat SARS-CoV-1 and related viruses, ultimately safeguarding human health. The effect of SARS-CoV-1 on human health can be lessened by developing prospective therapeutics and preventative measures by targeting key genes involved in viral replication, immune evasion, and cellular manipulation.

Figure 5 shows the heat map of first clustered genes. Each row in the heat map corresponds to a specific gene, and each column represents a sample or condition in which the gene expression was measured. The color intensity in each cell of the heat map indicates the level of gene expression, with brighter colors indicating higher expression and darker colors representing lower expression. Since the second cluster contains a diverse set of genes involved in various cellular functions, the heat map shows patterns of gene expression across different samples or conditions. Genes with similar expression patterns has been grouped together within the heat map. For example, ribosomal proteins show a distinct pattern of high expression across multiple samples, while other genes involved in RNA processing or splicing exhibit different expression patterns.

The second gene cluster was found by the study to be connected to several biological activities. Several ribosomal proteins found in it, such as RPL35, RPL28, RPL11, RPL21, RPL24, and RPS27A, are crucial for protein synthesis and have a significant impact on metabolism and cell function [9, 15, 16]. Additionally, SNORA27, SNORD102, SNORD14B, SNORD16, SNORD18A, SNORD18B, and SNORD18C, small nucleolar RNAs involved in RNA processing and modification. This suggests that active regulation of gene expression occurs in response to environmental factors, which may include viral infections like SARS-CoV-1 [17]. The presence of genes encoding RNA-binding proteins and translation initiation factors including EIF3F, EIF3D, and TARDBP in this cluster further suggests possible interactions with viral genetic material during translation [18]. The genes PSMB2, RNPS1, and DDX5, which are involved in RNA processing, splicing, and mRNA export, respectively [19, 20]. Viral infections may alter these processes, which are essential for gene expression and may have an effect on the cellular functions of the host. While Figure 6 displays the heat map of the second gene cluster. However, more experimental research is needed to confirm these genes' functional importance during viral infection in order to establish a clear and precise validation between these genes and SARS-CoV-1. In order to evaluate the clustered genes, the Silhouette Score has been used. The silhouette score is a measure used to assess the quality of clustering results [21]. It quantifies how well each gene that will fit into its assigned cluster compared to other clusters. For the given silhouette score of 0.5899, it suggests that the clustered genes are reasonably well-grouped, and the clustering has some level of separation between clusters.

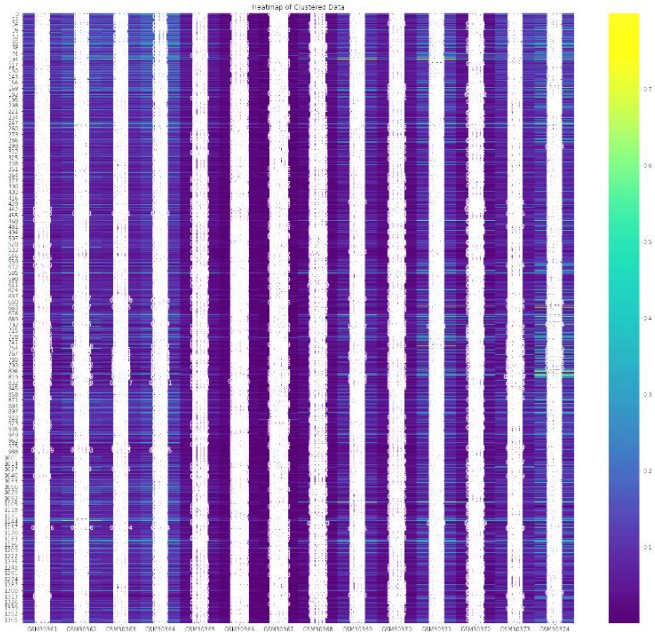

Fig. 5. Heat map of clustered genes-1

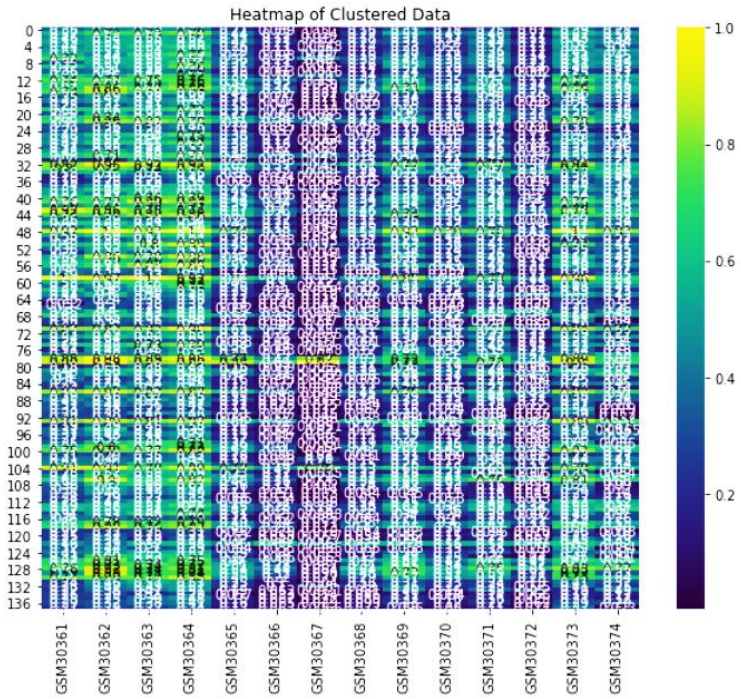

Fig. 6. Heat map of clustered genes-2

4. Conclusions

The study identified distinct clusters of genes with similar expression patterns, shedding light on the host's response to SARS-CoV-1 infection. Key genes involved in viral replication, immune response, and host-pathogen interactions showed significant alterations in expression levels, indicating their importance in the viral pathogenesis. The presence of subgroups within the infected samples suggests potential variations in the host response or viral strain differences. Overall, the research contributes to a comprehensive understanding of the virus-host interaction during the SARS-CoV-1 outbreak, and the identified transcriptional patterns may hold the potential for identifying therapeutic targets. Additionally, SARS-CoV-1 alters the cell cycle, DNA repair processes, and causes genomic instability, all of which may aid in the emergence of illnesses like cancer. For the purpose of developing effective ways to fight SARS-CoV-1 and similar viruses such as SARS-CoV-2, eventually protecting human health, a thorough knowledge of these disturbances is essential. Looking ahead, the utilization of clustering algorithms, akin to the approach outlined in this research [22], holds promise for the development of frameworks tailored to unsupervised technique in gene detection. Additional investigation and study of the discovered genes, including DDR1, MIR4640, RFC2, HSPA6, PAX8, and others, can help in the development of specific medications for viral infections and related disorders and offer useful insights into possible therapeutic targets and can be implemented within the framework.

Acknowledgement

This research work is supported by a grant from Xiamen University Malaysia Research Fund (XMUMRF/2022-C10/IECE/0037).

References

[1] Liu, Hsin-Liang, I-Jeng Yeh, Nam Nhut Phan, Yen-Hung Wu, Meng-Chi Yen, Jui-Hsiang Hung, Chung-Chieh Chiao et al. "Gene signatures of SARS-CoV/SARS-CoV-2-infected ferret lungs in short-and long-term models." *Infection, Genetics and Evolution* 85 (2020): 104438. https://doi.org/10.1016/j.meegid.2020.104438

[2] Thair, Simone A., Yudong D. He, Yehudit Hasin-Brumshtein, Suraj Sakaram, Rushika Pandya, Jiaying Toh, David Rawling et al. "Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections." *Iscience* 24, no. 1 (2021). https://doi.org/10.1016/j.isci.2020.101947

[3] Ramesh, Priyanka, Shanthi Veerappapillai, and Ramanathan Karuppasamy. "Gene expression profiling of corona virus microarray datasets to identify crucial targets in COVID-19 patients." *Gene Reports* 22 (2021): 100980. https://doi.org/10.1016/j.genrep.2020.100980

[4] Hamidi, Seyed Saeed, Ebrahim Akbari, and Homayun Motameni. "Consensus clustering algorithm based on the automatic partitioning similarity graph." *Data & Knowledge Engineering* 124 (2019): 101754. https://doi.org/10.1016/j.datak.2019.101754

[5] Oyelade, Jelili, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, and Ezekiel Adebiyi. "Clustering algorithms: their application to gene expression data." *Bioinformatics and Biology insights* 10 (2016): BBI-S38316. https://doi.org/10.4137/BBI.S38316

[6] Maulding, Nathan D., Spencer Seiler, Alexander Pearson, Nicholas Kreusser, and Joshua M. Stuart. "Dual RNA-Seq analysis of SARS-CoV-2 correlates specific human transcriptional response pathways directly to viral expression." *Scientific reports* 12, no. 1 (2022): 1329. https://doi.org/10.1038/s41598-022-05342-4

[7] Li, Teng, Amin Rezaeipanah, and ElSayed M. Tag El Din. "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement." *Journal of King Saud University-Computer and Information Sciences* 34, no. 6 (2022): 3828-3842. https://doi.org/10.1016/j.jksuci.2022.04.010

[8] Ruha, Leena, Tero Lähderanta, Lauri Lovén, Teemu Leppänen, Jukka Riekki, and Mikko J. Sillanpää. "Capacitated spatial clustering with multiple constraints and attributes." *arXiv preprint arXiv:2010.06333* (2021). https://doi.org/10.48550/arXiv.2010.06333

[9] Miller, Jessica, Nicole P. Hachmann, Ai-ris Y. Collier, Ninaad Lasrado, Camille R. Mazurek, Robert C. Patio, Olivia Powers et al. "Substantial neutralization escape by SARS-CoV-2 Omicron variants BQ. 1.1 and XBB. 1." *New England Journal of Medicine* 388, no. 7 (2023): 662-664. https://doi.org/10.1056/NEJMc2214314

[10] Jiang, Yanxialei, Jeeyoung Lee, Jung Hoon Lee, Joon Won Lee, Ji Hyeon Kim, Won Hoon Choi, Young Dong Yoo et al. "The arginylation branch of the N-end rule pathway positively regulates cellular autophagic flux and clearance of proteotoxic proteins." *Autophagy* 12, no. 11 (2016): 2197-2212. https://doi.org/10.1080/15548627.2016.1222991

[11] Dakal, Tikam Chand. "SARS-CoV-2 attachment to host cells is possibly mediated via RGD-integrin interaction in a calcium-dependent manner and suggests pulmonary EDTA chelation therapy as a novel treatment for COVID 19." *Immunobiology* 226, no. 1 (2021): 152021. https://doi.org/10.1016/j.imbio.2020.152021

[12] Abbas, Tarek, Mignon A. Keaton, and Anindya Dutta. "Genomic instability in cancer." *Cold Spring Harbor perspectives in biology* 5, no. 3 (2013): a012914. https://doi.org/10.1101/cshperspect.a012914

[13] Alexandrova, Elena, Giovanni Pecoraro, Assunta Sellitto, Viola Melone, Carlo Ferravante, Teresa Rocco, Anna Guacci et al. "An overview of candidate therapeutic target genes in ovarian cancer." *Cancers* 12, no. 6 (2020): 1470. https://doi.org/10.3390/cancers12061470

[14] Mughal, Muhammad Jameel, Kin Iong Chan, Ravikiran Mahadevappa, Sin Wa Wong, Kit Cheng Wai, and Hang Fai Kwok. "Over-activation of minichromosome maintenance protein 10 promotes genomic instability in early stages of breast cancer." *International Journal of Biological Sciences* 18, no. 9 (2022): 3827. https://doi.org/10.7150%2Fijbs.69344

[15] Wei, Jiajie, Rigel J. Kishton, Matthew Angel, Crystal S. Conn, Nicole Dalla-Venezia, Virginie Marcel, Anne Vincent et al. "Ribosomal proteins regulate MHC class I peptide generation for immunosurveillance." *Molecular cell* 73, no. 6 (2019): 1162-1173. https://doi.org/10.1016/j.molcel.2018.12.020

[16] El Khoury, Wendy, and Zeina Nasr. "Deregulation of ribosomal proteins in human cancers." *Bioscience Reports* 41, no. 12 (2021): BSR20211577. https://doi.org/10.1042/BSR20211577

[17] Wong, Yung-Hao, Cheng-Wei Li, and Bor-Sen Chen. "Evolution of network biomarkers from early to late stage bladder cancer samples." *BioMed Research International* 2014, no. 1 (2014): 159078. https://doi.org/10.1155/2014/159078

[18] Flynn, Ryan A., Julia A. Belk, Yanyan Qi, Yuki Yasumoto, Cameron O. Schmitz, Maxwell R. Mumbach, Aditi Limaye et al. "Systematic discovery and functional interrogation of SARS-CoV-2 viral RNA-host protein interactions during infection." *Biorxiv* (2020). https://doi.org/10.1101%2F2020.10.06.327445

[19] Latysheva, Natasha S., Matt E. Oates, Louis Maddox, Tilman Flock, Julian Gough, Marija Buljan, Robert J. Weatheritt, and M. Madan Babu. "Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer." *Molecular cell* 63, no. 4 (2016): 579-592. https://doi.org/10.1016/j.molcel.2016.07.008

[20] Legrand, Julien MD, Ai-Leen Chan, Hue M. La, Fernando J. Rossello, Minna-Liisa Änkö, Frances V. Fuller-Pace, and Robin M. Hobbs. "DDX5 plays essential transcriptional and post-transcriptional roles in the maintenance and

function of spermatogonia." *Nature communications* 10, no. 1 (2019): 2278. https://doi.org/10.1038/s41467-019-09972-7

[21] Gaido, Marco. "Distributed Silhouette Algorithm: Evaluating Clustering on Big Data." *arXiv preprint arXiv:2303.14102* (2023). https://doi.org/10.48550/arXiv.2303.14102

[22] Sobran, Nur Maisarah Mohd, and Zool Hilmi Ismail. "A systematic literature review of unsupervised fault detection approach for complex engineering system." *Journal of Advanced Research in Applied Mechanics* 103, no. 1 (2023): 43-60. https://doi.org/10.37934/aram.103.1.4360