Differential Expression Analysis of Longitudinal Study of Septic and Cardiogenic Shock Using Limma

Chandana Prakash and Shreyoshi Ghosh

Abstract

Septic shock and cardiogenic shock are common types of circulatory shock, the clinical expression of circulatory failure. Shock is prevalent in the ICU, affecting about one third of the patients. While the molecular mechanisms associated with septic shock have been widely studied, factors involved in cardiogenic shock have not. The goal of this project was to take a different approach than the paper when looking at the differentially expressed genes involved in both types of shock to understand the common GO terms and pathways associated with each. A limma analysis was performed to study the time contrasts two ways: a) within shock groups and b) between shock groups. To do this we used the count data obtained from Salmon, which was mapped to the human transcriptome, and the count data provided by the paper, which was mapped to the human genome. Genes with p value < 0.05 (Benjamini-Hochberg multiple test correction) were defined differentially expressed (DEGs). We observed a difference in the log fold change, which impacted the number of up and down regulated genes found between both count data sets. Overrepresentation analysis was used to identify the biological processes significantly enriched in both types of shock. We found GO terms associated with the innate immune response, cell cycle and DNA repair pathways to be most commonly enriched and the neutrophil degranulation pathway to have the most hits in both count datasets. Further analysis with a larger data set would aid in overcoming the present obstacle of studying the direction of regulation of the enriched pathways.

Introduction

Circulatory shock can be defined as the clinical expression of circulatory failure that results in inadequate cellular oxygen utilization [1, 2]. Two common pathophysiological mechanisms that cause circulatory shock are distributive factors and cardiogenic factors, which respectively cause septic shock and cardiogenic shock. Shock is extremely prevalent in the ICU, affecting about one third of the patients, with septic shock being the most common [3].

Septic shock (SS) is the most severe complication of sepsis, which is a whole body immune response to an infection. In septic shock, multiple organ system failure begins along with a dramatic drop in blood pressure. Cardiogenic shock (CS) is rarer, and is typically the result of a severe heart attack. It occurs when the heart is unable to pump enough blood or oxygen to the brain and other organs. Both types of shocks have different etiology but result in similar consequences and mortality rates, with SS having a rate of 30% and CS 40% [1].

While the molecular mechanisms associated with SS have been widely explored, factors involved in CS have been very poorly studied. Braga, D. et al studied the RNA sequencing results to try to explore the transcriptome in whole blood of septic and cardiogenic shock patients at different time points of ICU stay [1]. The goal of our project was to take a different approach from the paper when looking at the differentially expressed genes involved in both types of shock to understand the common GO terms and pathways associated with each. We applied the same approach to perform a comparative analysis between the count data we produced and the count data provided by the paper.

The data used in the referenced paper was part of the multicenter prospective observational trial ShockOmics, and consisted of whole blood samples from adult ICU patients from Geneva and Belgium [12]. The blood samples were collected at three time points: T1,

within 16 h of ICU admission; T2, 48 h after study enrollment; and T3, on day 7 from ICU admission or before discharge from the ICU [1].

Methods

Raw Reads Quality Check and Preprocessing

We chose 6 samples from each type of shock, with our exclusion criteria being male and no mortality. The raw reads were obtained from the GEO database (accession number GSE131411). A quality check was performed on these samples using FastQC and MultiQC in order to analyze the quality of the reads [7]. We chose not to preprocess the data because of the overall quality statistics.

Sequencing Data Analysis

We used quasi-mapping to quantify the paired-end reads against the human transcript sequences from GENCODE as reference using Salmon (version 1.4.0). Tximport from BioConductor was used to create a count matrix and assign the reads to genes [9, 10].

The counts table generated by Braga D et al. was generated by aligning raw reads against human reference genome (GRCh38) using STAR alignment tool and featureCounts [1].

Differential Expression Analysis

DEA was done in order to classify differentially expressed genes (DEGs) with similar expression patterns. Built-in functions from the BioConductor package edgeR were used for data preprocessing. Voom was then used to transform the data prior to linear modeling in Limma. Time series analysis was performed using Limma to study the gene expression changes over time in CS and SS patients with different paired analyses comparing 1) the same time point across the SS and CS, and 2) T1 to T3 in both CS and SS separately. Genes with padj < 0.05 and Benjamini-Hochberg multiple test correction for FDR were considered differentially expressed

and used for downstream analysis. Before other downstream analyses, clustering analyses such as hierarchical clustering and principal component analyses (PCA) were performed to ensure that the samples were clustered together based on the experimental design, such as by biological replicates and not by technical factors such as number of reads sequenced, etc. [1, 7].

GOterm and Pathway Enrichment Analysis

We used BioMart to fetch the EntrezIDs used in the GO term analysis. We performed a GO term enrichment analysis to identify enriched biological processes in both types of shocks using GOstats, and GOenrich was used to visualize the enriched GO terms [11]. The Reactome database was used to visualize and interpret enriched pathways for the differentially expressed genes [4].

Results

Quality check

The overall statistics and mean quality scores (Fig 1) suggests that the raw reads are of good quality. We did not rid the duplicates so as to not limit our total sample size for the analysis.

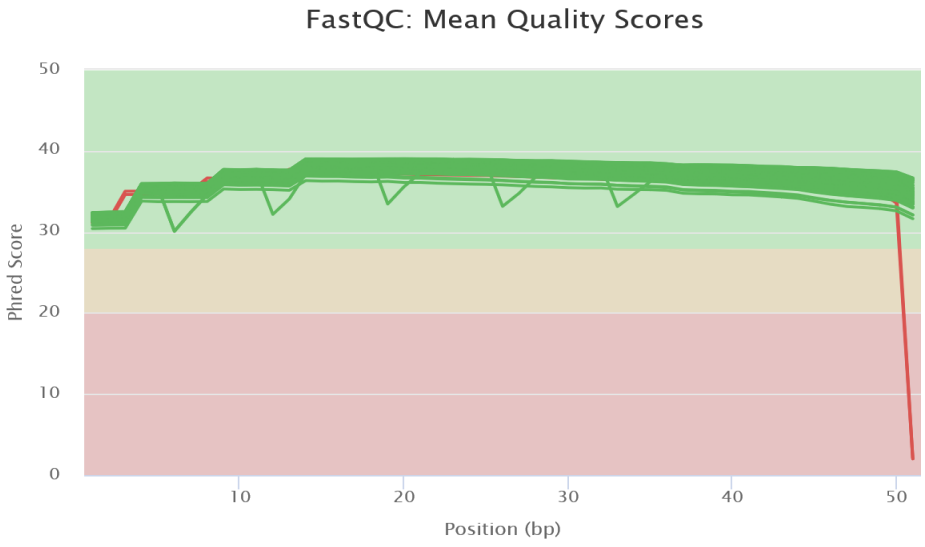
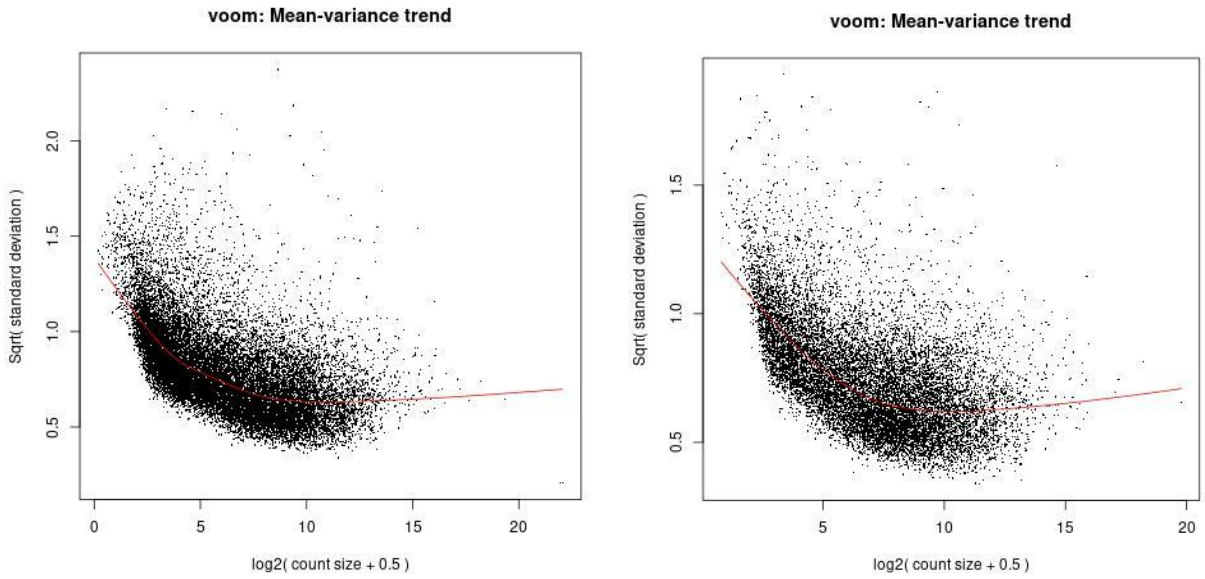


Fig 1: FastQC per base sequence mean quality score plot

Voom

Voom modifies RNA-Seq data for use with limma and shows how the coefficient of variation of the counts depends on the count size. In our case, it shows a decreasing trend between the means and variances resulting from a combination of technical and biological variations (Fig 2).

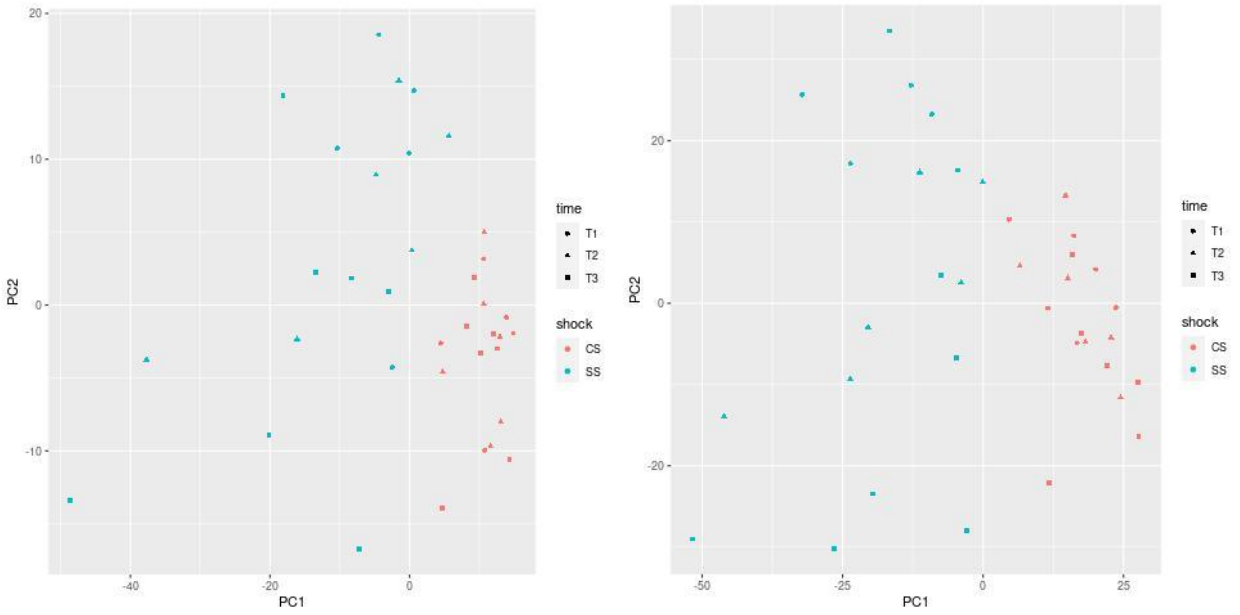


(a) (b)

Figure 2: Voom mean-variance trend for (a) our count data and (b) the paper's count data

PCA

We performed PCA on the most DEGs across samples between shock groups. It was seen that the PCA results (Fig 3) were grouped by shock type and then within each shock type, SS had a clear distinction between the three time points while CS had no clear organization regarding the time points.

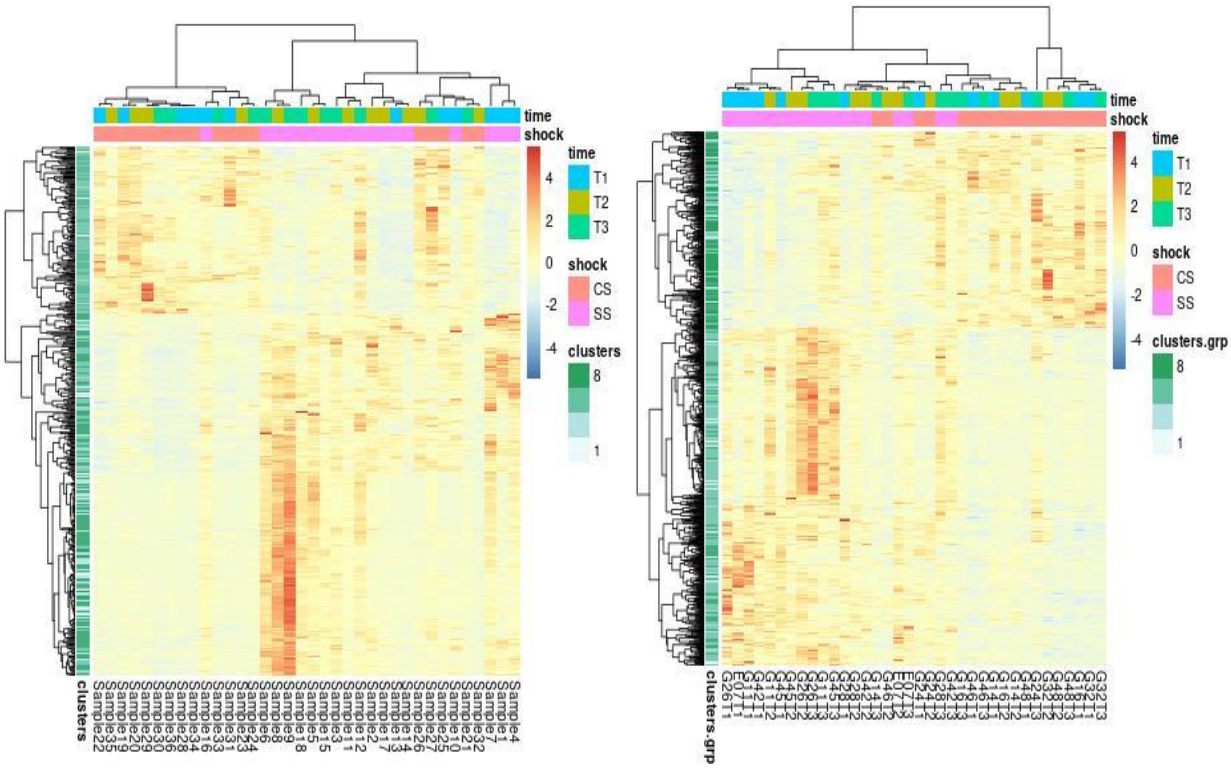


(a) (b)

Figure 3: PCA graph for (a) our count data and (b) the paper's count data

Heatmap

Hierarchical clustering with Pearson's correlation was performed to visualize differentially expressed genes between shock groups using a heatmap (Fig 4). We observed a large number of highly expressed significant genes within the septic shock group, while there were a lot of lowly expressed significant genes within the cardiogenic group across different time points in data from both count tables.



(a) (b)

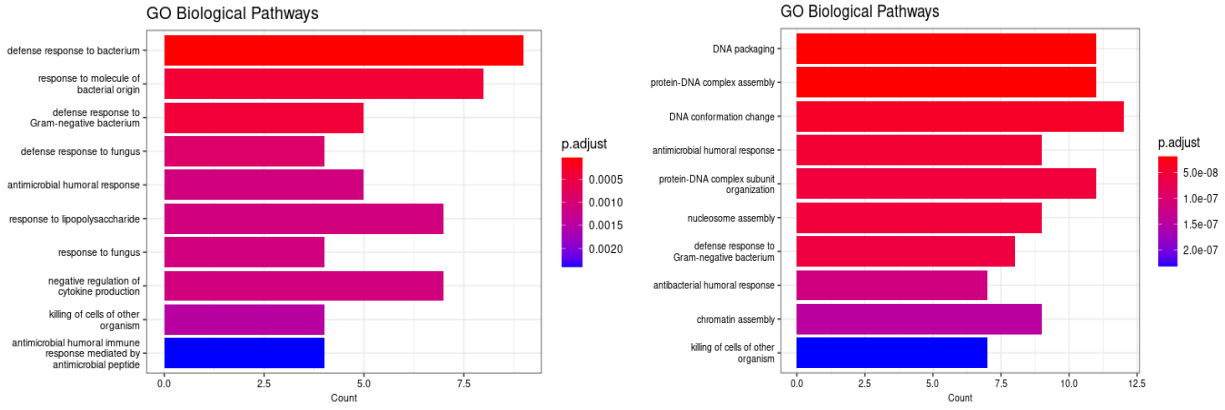
Figure 4: Heatmap for (a) our count data and (b) the paper's count data

Gene Set Enrichment Analysis

When looking at how many genes were up or down regulated, our count data showed 8 up and 11 down regulated genes in T1, 2 up and 1 down regulated genes in T2, and 0 up and 28 down regulated genes in T3. The count data provided by the paper showed 316 up and 409 down regulated genes in T1, 22 up and 84 down regulated genes in T2, and 4 up and 96 down regulated genes in T3. The enriched GO terms we obtained were for overrepresented genes, however without the distinction of up or down regulation.

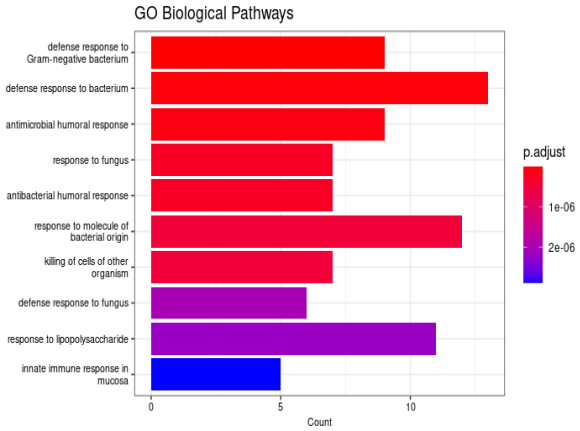
Bar Plot

We created bar plots for each time point to visualize the count and enrichment level of the GO pathways [11]. It was seen that in both time 1 (Fig 5a) and time 3 (Fig 5c) the defense response to bacterium pathway is the most enriched with the highest gene count. For time 2 (Fig 5b), DNA conformation change is the most enriched.



(a)

(b)

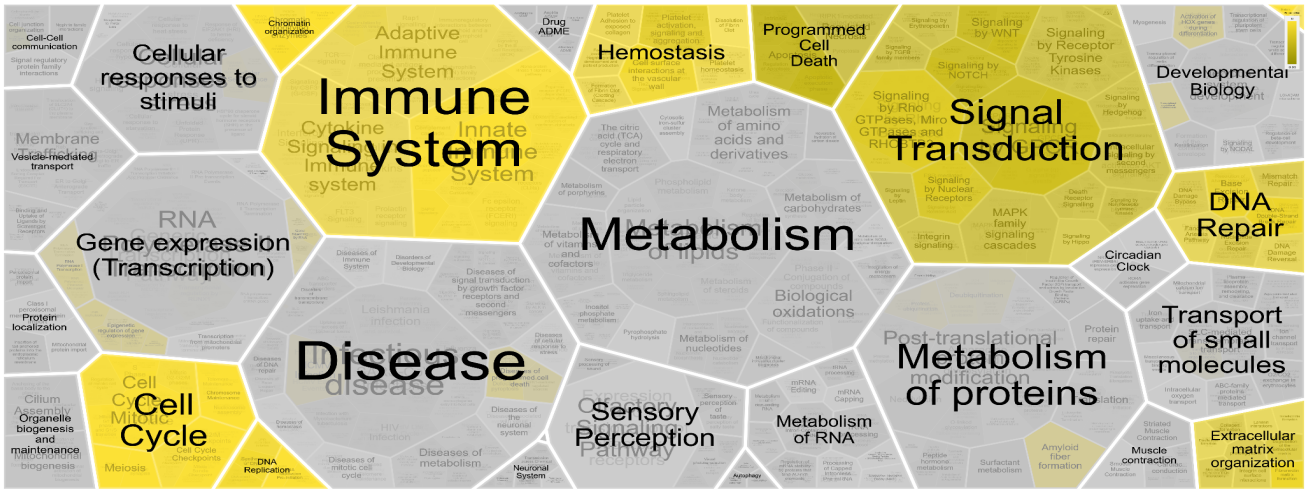


(c)

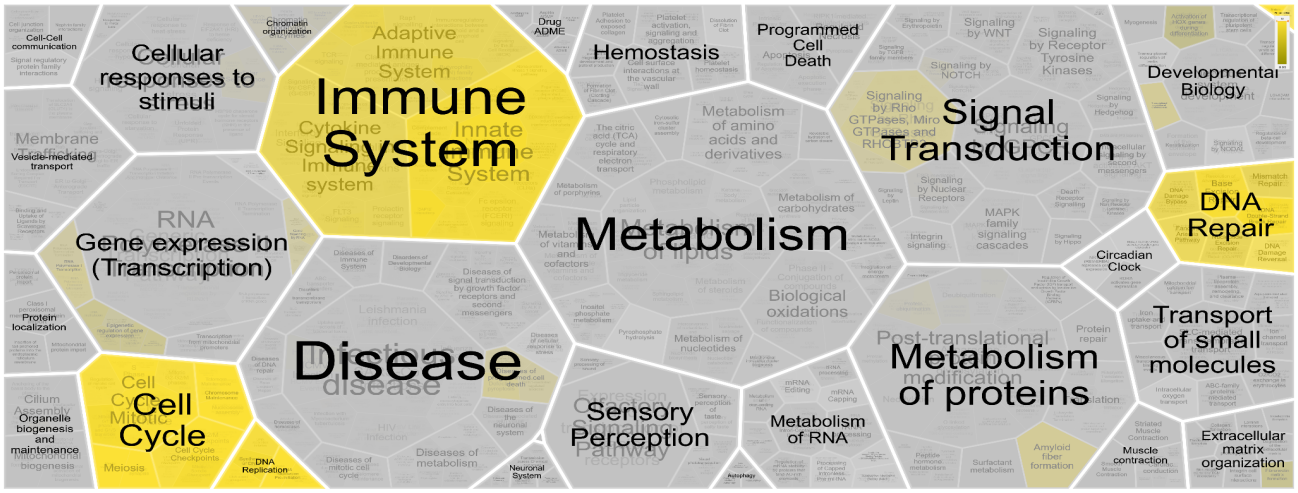
Figure 5: Bar Plot for (a) Time Point 1 (b) Time Point 2 and (c) Time Point 3

Reactome

The EntrezIDs for the DEGs from both count data sets were submitted to the Reactome database to look at the most enriched pathways (over-represented). For our data, 305 out of 385 identifiers in the sample were found in Reactome, where 1136 pathways were hit by at least one of them (Fig 6a). For the paper's count data, 744 out of 1027 identifiers in the sample were found in Reactome, where 1645 pathways were hit by at least one of them (Fig 6b).



(a)



(b)

Figure 6: Reactome Results for (a) our count data and (b) the paper's count data

We obtained the top 25 pathways, of which neutrophil degranulation had the most hits with both count matrices (Fig 7).

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Neutrophil degranulation	59 / 480	0.032	1.11e-16	1.35e-13	10 / 10	7.25e-04

(a)

Pathway name	Entities				Reactions	
	found	ratio	p-value	FDR*	found	ratio
Neutrophil degranulation	96 / 480	0.022	1.11e-16	1.96e-13	10 / 10	7.25e-04

(b)

Figure 7: Most relevant pathway of (a) our count data and (b) the paper's count data

Discussion

In our analysis using Salmon count data, within-group transcriptomic analysis over time showed no significant modulation in both shock groups. When using count data provided by the paper, we observed 581 up and 752 down genes within the septic shock group and none within the cardiogenic shock group. Analysis between-groups at the same time points showed significant modulation in 461 genes, with a larger number of genes differentially expressed at T3 compared to T1, while no genes were observed at T2. The same between-group design was studied with the count data generated by the paper and we observed a total of 1222 variable genes, with a large number of genes at T1 in comparison to the other time points. We found 385 genes to be common in both counts data sets. We observed a difference in the log fold change, which impacted the number of up and down regulated genes found between both count data sets.

The common genes and pathways in both of the shock groups were identified through the differential expression in the time frames. The overrepresentation analysis of GO enriched terms associated with biological processes such as innate immune response, response to stress/shock,

nucleosome assembly is representative of the processes generally involved in circulatory shock. It could be that a large portion of the GO enriched terms were associated specifically with septic shock.

The analysis using the Reactome database and the count data provided by the paper indicated that irrespective of etiology, we observed in both SS and CS groups common pathways of inflammation, DNA replication and immunoglobulins. This also confirms the GO enriched terms in the overrepresentation analysis. Along with these commonly modulated pathways , we also observed other significant pathways such as signal transduction and hemostasis using our own Salmon counts data. The signal transduction pathway is critical in identifying molecular targets for pharmacological intervention regarding cardiovascular disorders. This is because changes in signaling systems can cause differences in the functions of the cells in the heart and vascular wall [5].

The most relevant pathway found in the Reactome analysis is neutrophil degranulation. Neutrophils play an important role in regards to inflammation and immunity, which are critical in the body's defense, specifically, the production of inflammatory cytokines and the release of reactive oxygen species. The degranulation of the neutrophil granules kills pathogens. In recent studies, it has been seen that neutrophils are related to the severity of myocardial infarctions, a major cardiac event, potentially leading to cardiogenic shock [6].

A large limitation of this analysis was having a small sample size. This hindered the detection of the relevant gene expression changes and made it harder to find the differences between the phenotypic subgroups due to the small groups of patients. Additionally, sepsis being a heterogeneous condition resulted in a high variability between the patients [1, 8]. The paper also created its count data and performed the analysis with the inclusion of a third group, a

control no shock group. We were unable to obtain this data, causing a significant change in the analysis we were able to do.

In our comparative study of the count table obtained by transcriptome versus genome mapping [7], although there was not much difference in the number of genes in both count tables, we observed a huge difference in the differential expression analysis. At the moment we are unable to tell why the results from different count tables vary, and this is something that could be further investigated.

The possibility of shared mechanistic pathways between the types of shock could aid in creating personalized therapies for various illnesses and conditions by finding common targets [1]. In the future, the next step would be to perform a different study using a larger sample size in order to try to analyze the relationships between these different factors. It would also be useful to do a further analysis of the enriched pathways at different time points.

References

1. Braga D, Barcella M, Herpain A, *et al.* A longitudinal study highlights shared aspects of the transcriptomic response to cardiogenic and septic shock. *Crit Care* 23, 414 (2019). https://doi.org/10.1186/s13054-019-2670-8
2. Bates ER. *Cardiac Intensive Care (3rd Ed.)* Elsevier, (2019).
3. Vincent JL, De Backer D. Circulatory shock. N Engl J Med. 2013 Oct 31;369(18):1726-34. doi: 10.1056/NEJMra1208943. PMID: 24171518.
4. Fabregat, A., Sidiropoulos, K., Viteri, G. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 18, 142 (2017). https://doi.org/10.1186/s12859-017-1559-2
5. Wheeler-Jones CP. Cell signaling in the cardiovascular system: an overview. *Heart*. 2005;91(10):1366-1374. doi:10.1136/hrt.2005.072280
6. Zhang N, Aiyasiding X, Li WJ, Liao HH, Tang QZ. Neutrophil degranulation and myocardial infarction. Cell Commun Signal. 2022 Apr 11;20(1):50. doi: 10.1186/s12964-022-00824-4. PMID: 35410418; PMCID: PMC8996539.
7. Conesa A, Madrigal P, Tarazona S, *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**
8. Marshall JC. Why have clinical trials in sepsis failed? Trends Mol Med. 2014;20:195–203
9. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419. doi:10.1038/nmeth.4197
10. Love MI, Soneson C, Patro R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Res*. 2018;7:952. Published 2018 Jun 27. doi:10.12688/f1000research.15398.3
11. Mohammed Khalfan, *Over-Representation Analysis with ClusterProfiler.* NGS Analysis. https://learn.gencore.bio.nyu.edu/rna-seq-analysis/over-representation-analysis/
12. Aletti F, Conti C, Ferrario M, et al. ShockOmics: multiscale approach to the identification of molecular biomarkers in acute heart failure induced by shock. *Scand J Trauma Resusc Emerg Med*. 2016;24:9. Published 2016 Jan 28. doi:10.1186/s13049-016-0197-4