Replicating and Validating Results from "Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients"

Byron Zhang

Abstract

Septic shock, a serious health condition caused by uncontrolled sepsis, is particularly dangerous to infants and children. We replicated the study "Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients" and identified 65 differentially expressed genetic biomarkers within survivors and non-survivors of 181 septic shock patients admitted to the pediatric intensive care unit (PICU). Through performing functional enrichment analysis on these biomarkers, we identified several biological processes and pathways associated with septic shock development, including immune response, cell division, and chemokin activity. The identification of these biomarkers and their associated biological processces and pathways may help facilitate early diagnosis, and can support more effective treatment and prognosis. Although we were unable to recapitulate exactly the results presented in the original study, we were able to obtain partially similar results, uncovering some genes and pathways presented in the original study. Furthermore, we found that different background noise correction techniques for data preprocessing yielded different results in both differential expression analysis and functional term enrichment, which brings up the discussion of how to select the optimal noise correction procedure for future work.

1 Background

Pediatric septic shock is a life threatening organ dysfunction caused by an imbalanced host response to infection in children [1]. In recent years, the development of multi-omics sequencing technologies allowed for researchers to generate more genetic data, uncover more genetic biomarkers and perform more thorough analysis on roles of these biomarkers in human diseases. The study "Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients" by Mohammed et al. attempts to takes advantage of available gene expression data to identify diferentially expressed pediatric septic shock biomarkers between survivors and non-survivors, which may serve as predictors for long-term survival outcomes. The study first analyzed gene expression data from 181 blood samples of critically-ill patients within 24 hours of admission to the PICU. Then, functional enrichment analysis was performed to confirm previous pathways and functional groups related to septic shock and to identify new ones. The results of the study may provide new insights into septic shock diagnosis, prognosis, and treatment. [2]

The primary goals for our project are to: (1) reproduce Mohammed et al.'s original study, (2) verify the original study's results, (3) address any discrepancies between our results and the original study's results, **(4) compare the effects of different noise correction techniques not employed in the data preprocessing stage of the original study**.

2 Results & Figures

We replicated the study with different methods to correct for background noise on raw data, once using the Robust Multi-Array Analysis (RMA) algorithm [3], and once using the Guanine Cytosine Robust Multi-Array Analysis (gcRMA) algorithm [4]. For each noise correction method, genes with an adjusted p-value less than 0.05 using the Benjamini-Hochberg method [5] are considered differentially expressed, and their associated log fold

changes, average expressions, t-statistics, p-values, adjusted p-values are reported (See Tables 4, 5). Functional enrichment statistics for both noise correction methods with an adjusted p-value threshold of 0.05 are also reported below in Tables 6, 7. We also produce a heatmap illustrating the log expression values of the differentially expressed genes discovered using RMA-corrected data, with annotations such as immunosuppression and gender for each patient (See table 2). Result tables and figures from the original study is also presented below for comparison.

3 Discussion

3.1 Interpretation of Results

As shown by Table 4, by replicating the steps of the original study exactly, we were able to successfully uncover the 16 differentially expressed genes reported in Table 1 of the original study, as all those genes are found to be significant in our study with BH-adjusted p-value below the 0.05 threshold. However, the log2 fold change, average expression, t-statistic, p-values and adjusted p-values of these 16 genes in our study differ from those of the original study. Furthermore, our analysis differed from the original study's in terms of the number of differentially expressed genes discovered. Our anlysis showed that there are 65 differentially expressed genes, with 55 upregulated and 10 downregulated, while the original study reported only 54, with 47 upregulated and 7 downregulated. We were unable to determine whether or not the 54 genes found in the original study is a subset of the 65 genes in our study, since the original study did not provide details on the identity of the 54 genes they uncovered.

In addition, we found that using different background noise correction techniques during the data preprocessing stage of the experiment yielded different results in terms of both differential gene analysis and functional term enrichment analysis. We observed that the differentially expressed genes we uncovered using gcRMA are very different from those that

Etiology	Number of patients	Total (%)
Streptococcus pneumoniae	17	9.39
Staphylococcus aureus	17	9.39
Streptococcus pyogenes	12	6.63
Klebsiella	9	4.97
Streptococcus agalactiae	7	3.87
Neisseria meningitidis	7	3.87
Enterococcus	5	2.76
Gram-negative rods	4	2.21
Simplexvirus	3	1.66
influenza A virus	3	1.66
Mixed	3	1.66
Enterobacter	2	1.11
Hemophilus	2	1.11
Pseudomonas	2	1.11
Serratia	2	1.11
Acinetobacter	1	0.55
Adenovirus	1	0.55
BK virus	1	0.55
Candida	1	0.55
Clostridium	1	0.55
Cytomegalovirus	1	0.55
Escherichia coli	1	0.55
Group F Streptococcus	1	0.55
Human metapneumovirus	1	0.55
Moraxella	1	0.55
Parainfluenza	1	0.55
Rickettsia	1	0.55
Unknown	74	40.88

Table 1: Etiology table from the original study, listing the organism involved in the development of septic shock of the 181 patients used in the study.

Etiology	Frequency	Percentage
None	74	40.8839779
Staph aureus	17	9.3922652
Pneumococcus	17	9.3922652
Group A Strep	12	6.6298343
Klebsiella	9	4.9723757
Meningococcus	7	3.8674033
Group B Strep	7	3.8674033
Enterococcus	5	2.7624309
Gram negative rods	4	2.2099448
Mixed (Klebsiella + Enterococcus + Candida)	3	1.6574586
Influenza A	3	1.6574586
Herpes Simplex Virus	3	1.6574586
Serratia	2	1.1049724
Pseudomonas	2	1.1049724
Hemophilus	2	1.1049724
Enterobacter	2	1.1049724
Rickettsia	1	0.5524862
Parainfluenza	1	0.5524862
Moraxella	1	0.5524862
Human metapneumovirus	1	0.5524862
Group F Strep	1	0.5524862
E coli	1	0.5524862
Cytomegalovirus	1	0.5524862
Clostridium	1	0.5524862
Candida	1	0.5524862
BK virus	1	0.5524862
Adenovirus	1	0.5524862
Acinetobacter	1	0.5524862

Table 2: Our replication of Table 1 (BZ and JYo). The "frequencies" column here corresponds to the number of patients in Table 1 and should be exactly the same as Table 1.

Gene	Fold Change	Average Expression	t-statistics	p-value	adj. p-value	Reference
DDIT4	2.12	8.441545	5.490723	1.33E-07	0.000592	[26,49]
CCL3	2.12	6.208106	4.627029	7.02E-06	0.012717	[29]
PRG2	2.11	5.48492	5.611777	7.35E-08	0.000533	—
MT1M	1.78	3.72997	5.641576	6.35E-08	0.000533	—
CDC20	1.68	6.18643	4.059008	7.31E-05	0.040893	[25]
KIF20A	1.66	4.940242	4.673723	5.74E-06	0.012717	—
MAFF	1.64	6.815799	4.437113	1.58E-05	0.015559	[50]
EBI3	1.64	5.41593	4.517058	1.12E-05	0.013575	[51]
MELK	1.63	6.413027	4.141523	5.27E-05	0.034718	—
TOP2A	1.58	4.997892	4.045113	7.72E-05	0.040893	[52]
NUSAP1	1.54	6.627514	3.928091	0.000121	0.049772	[53]
RGL1	1.52	7.288761	4.447937	1.51E-05	0.015559	[54]
ARHGEF40	−1.66	6.880613	−3.994983	9.38E-05	0.043368	—
LOC254896	−1.65	8.498379	−4.452352	1.48E-05	0.015558	—
SLC46A2	−1.61	5.880150	−4.084894	6.60E-05	0.039833	—
TNFRSF10C	−1.54	8.678817	−4.643272	6.55E-06	0.012717	[55]

Table 3: 16 of the 54 differentially expressed genes found in the original study with the largest absolute log fold change. This table comes from the original study.

Gene	logFC	AverExpr	t-sttatistic	P-Value	adj.P.Val
DDIT4	1.084952	8.441548	5.491076	1.33E-07	0.0005691
CCL3	1.083665	6.208105	4.62732	7.01E-06	0.0122433
PRG2	1.079349	5.484923	5.612079	7.34E-08	0.0005126
MT1M	0.828031	3.729966	5.641984	6.33E-08	0.0005126
CDC20	0.752058	6.18643	4.059279	7.30E-05	0.0384393
KIF20A	0.727482	4.940237	4.674027	5.73E-06	0.0122433
MAFF	0.714831	6.8158	4.437449	1.57E-05	0.014977
EBI3	0.709876	5.41593	4.517347	1.12E-05	0.0130695
MELK	0.701697	6.41303	4.141809	5.27E-05	0.0324461
TOP2A	0.661859	4.997893	4.045393	7.71E-05	0.0384393
CTSL	0.636553	8.251058	3.873242	1.50E-04	0.0489695
ORM2	0.636438	5.943288	4.38968	1.92E-05	0.0174862
ANLN	0.627523	4.06938	3.918639	1.26E-04	0.0468674
NUSAP1	0.618912	6.627514	3.928403	1.21E-04	0.0468674
RGL1	0.6056	7.288761	4.448407	1.50E-05	0.014977
H1-0	0.602908	7.112091	3.885217	1.43E-04	0.0483034
TCEAL9	0.553585	4.039555	4.672907	5.76E-06	0.0122433
SDC4	0.527194	6.026719	5.486261	1.36E-07	0.0005691
CCL20	0.51599	3.919642	3.955444	1.09E-04	0.0438944
CDK1	0.513161	4.991024	4.094776	6.35E-05	0.0369296
HJURP	0.50299	5.422606	4.562932	9.25E-06	0.0130695
TUBB	0.501598	9.595086	3.901988	1.34E-04	0.0468674
KIF2C	0.47683	5.888558	4.669751	5.84E-06	0.0122433
CCL2	0.473976	4.03389	4.340539	2.35E-05	0.0197168
SLCO4A1	0.467039	5.634885	4.967542	1.55E-06	0.0054265
SLC39A8	0.457006	5.436631	4.574643	8.80E-06	0.0130695
PRG3	0.439296	6.460666	4.448225	1.50E-05	0.014977
NDUFV2	0.437183	8.431119	4.033734	8.07E-05	0.039297
NEK2	0.423943	4.816173	4.23499	3.62E-05	0.0250073
EPDR1	0.423108	4.737912	4.32102	2.55E-05	0.0205446
ID1	0.412429	4.429473	3.910242	1.30E-04	0.0468674
JUN	0.407501	5.028367	5.912444	1.63E-08	0.0003414
CDC45	0.388105	5.828917	4.055913	7.40E-05	0.0384393
TFRC	0.378655	6.525043	3.975181	1.01E-04	0.0424299
LMNA	0.376462	6.109634	4.209946	4.01E-05	0.0262356
SPAG5	0.346713	6.125419	3.95127	1.11E-04	0.0438944
CFHR1	0.338406	3.670181	3.95769	1.08E-04	0.0438944
ESPL1	0.334885	5.976305	4.343374	2.33E-05	0.0197168
EMP1	0.329284	4.06196	4.262172	3.24E-05	0.0234302
AURKB	0.313932	5.428148	3.880772	1.45E-04	0.0483459
HMGB3P1	0.309345	6.140289	4.53568	1.04E-05	0.0130695
APOC1	0.308169	6.122925	4.629476	6.95E-06	0.0122433
RAD51	0.308098	4.655542	4.075517	6.85E-05	0.0377467
BTG3	0.305495	4.882596	4.301056	2.77E-05	0.0214722
CXCL3	0.296013	4.409862	4.229678	3.70E-05	0.0250073
C11orf96	0.293868	4.452529	4.526786	1.08E-05	0.0130695
CCNB2	0.286718	5.488862	4.175904	4.60E-05	0.0291715
AREG	0.271759	3.375459	4.278093	3.04E-05	0.022743
FSCN1	0.265836	4.859574	3.852853	1.62E-04	0.0519503
GTSE1	0.263302	6.111296	3.907404	1.31E-04	0.0468674
REXO5	0.233642	4.7662	4.129763	5.53E-05	0.0330669
ETV5	0.216336	4.658393	4.050669	7.55E-05	0.0384393
SCD	0.215349	6.021515	4.020555	8.49E-05	0.0404306
SPRED2	0.208714	4.20909	4.534057	1.05E-05	0.0130695
CCNF	0.207789	5.17316	4.000191	9.19E-05	0.0408941
RCBTB2	−0.37629	6.47148	−4.008519	8.90E-05	0.040528
LOC401261	−0.42693	6.554133	−3.897425	1.36E-04	0.0468674
NAAA	−0.44176	7.053133	−3.901223	1.35E-04	0.0468674
NUAK2	−0.56783	9.340512	−4.012585	8.76E-05	0.040528
CD86	−0.57443	7.043634	−3.987156	9.67E-05	0.0413356
TLR6	−0.61355	7.071961	−3.903954	1.33E-04	0.0468674
TNFRSF10C	−0.62231	8.678819	−4.643669	6.54E-06	0.0122433
SLC46A2	−0.68594	5.88015	−4.085178	6.59E-05	0.0373188
LOC254896	−0.71949	8.498379	−4.452628	1.48E-05	0.014977
ARHGEF40	−0.73174	6.880613	−3.995246	9.37E-05	0.0408941

Table 4: All 65 differentially expressed genes discovered with RMA noise correction in our analysis (JYo). Here we see that the genes at the beginning and the end of the list are the ones with the largest absolute log fold change. Observe that all genes in Table 3 are part of this table.

Gene	logFC	AverExpr	t-sttatistic	P-Value	adj.P.Val
DDIT4	1.084952	8.441548	5.491076	1.33E-07	0.0005691
CCL3	1.083665	6.208105	4.62732	7.01E-06	0.0122433
PRG2	1.079349	5.484923	5.612079	7.34E-08	0.0005126
MT1M	0.828031	3.729966	5.641984	6.33E-08	0.0005126
CDC20	0.752058	6.18643	4.059279	7.30E-05	0.0384393
KIF20A	0.727482	4.940237	4.674027	5.73E-06	0.0122433
MAFF	0.714831	6.8158	4.437449	1.57E-05	0.014977
EBI3	0.709876	5.41593	4.517347	1.12E-05	0.0130695
MELK	0.701697	6.41303	4.141809	5.27E-05	0.0324461
TOP2A	0.661859	4.997893	4.045393	7.71E-05	0.0384393
CTSL	0.636553	8.251058	3.873242	1.50E-04	0.0489695
ORM2	0.636438	5.943288	4.38968	1.92E-05	0.0174862
ANLN	0.627523	4.06938	3.918639	1.26E-04	0.0468674
NUSAP1	0.618912	6.627514	3.928403	1.21E-04	0.0468674
RGL1	0.6056	7.288761	4.448407	1.50E-05	0.014977
H1-0	0.602908	7.112091	3.885217	1.43E-04	0.0483034
TCEAL9	0.553585	4.039555	4.672907	5.76E-06	0.0122433
SDC4	0.527194	6.026719	5.486261	1.36E-07	0.0005691
CCL20	0.51599	3.919642	3.955444	1.09E-04	0.0438944
CDK1	0.513161	4.991024	4.094776	6.35E-05	0.0369296
HJURP	0.50299	5.422606	4.562932	9.25E-06	0.0130695
TUBB	0.501598	9.595086	3.901988	1.34E-04	0.0468674
KIF2C	0.47683	5.888558	4.669751	5.84E-06	0.0122433
CCL2	0.473976	4.03389	4.340539	2.35E-05	0.0197168
SLCO4A1	0.467039	5.634885	4.967542	1.55E-06	0.0054265
SLC39A8	0.457006	5.436631	4.574643	8.80E-06	0.0130695
PRG3	0.439296	6.460666	4.448225	1.50E-05	0.014977
NDUFV2	0.437183	8.431119	4.033734	8.07E-05	0.039297
NEK2	0.423943	4.816173	4.23499	3.62E-05	0.0250073
EPDR1	0.423108	4.737912	4.32102	2.55E-05	0.0205446
ID1	0.412429	4.429473	3.910242	1.30E-04	0.0468674
JUN	0.407501	5.028367	5.912444	1.63E-08	0.0003414
CDC45	0.388105	5.828917	4.055913	7.40E-05	0.0384393
TFRC	0.378655	6.525043	3.975181	1.01E-04	0.0424299
LMNA	0.376462	6.109634	4.209946	4.01E-05	0.0262356
SPAG5	0.346713	6.125419	3.95127	1.11E-04	0.0438944
CFHR1	0.338406	3.670181	3.95769	1.08E-04	0.0438944
ESPL1	0.334885	5.976305	4.343374	2.33E-05	0.0197168
EMP1	0.329284	4.06196	4.262172	3.24E-05	0.0234302
AURKB	0.313932	5.428148	3.880772	1.45E-04	0.0483459
HMGB3P1	0.309345	6.140289	4.53568	1.04E-05	0.0130695
APOC1	0.308169	6.122925	4.629476	6.95E-06	0.0122433
RAD51	0.308098	4.655542	4.075517	6.85E-05	0.0377467
BTG3	0.305495	4.882596	4.301056	2.77E-05	0.0214722
CXCL3	0.296013	4.409862	4.229678	3.70E-05	0.0250073
C11orf96	0.293868	4.452529	4.526786	1.08E-05	0.0130695
CCNB2	0.286718	5.488862	4.175904	4.60E-05	0.0291715
AREG	0.271759	3.375459	4.278093	3.04E-05	0.022743
FSCN1	0.265836	4.859574	3.852853	1.62E-04	0.0519503
GTSE1	0.263302	6.111296	3.907404	1.31E-04	0.0468674
REXO5	0.233642	4.7662	4.129763	5.53E-05	0.0330669
ETV5	0.216336	4.658393	4.050669	7.55E-05	0.0384393
SCD	0.215349	6.021515	4.020555	8.49E-05	0.0404306
SPRED2	0.208714	4.20909	4.534057	1.05E-05	0.0130695
CCNF	0.207789	5.17316	4.000191	9.19E-05	0.0408941
RCBTB2	-0.37629	6.47148	-4.008519	8.90E-05	0.040528
LOC40126[illegible]	-0.42693	6.554133	-3.897425	1.36E-04	0.0468674
NAAA	-0.44176	7.053133	-3.901223	1.35E-04	0.0468674
NUAK2	-0.56783	9.340512	-4.012585	8.76E-05	0.040528
CD86	-0.57443	7.043634	-3.987156	9.67E-05	0.0413356
TLR6	-0.61355	7.071961	-3.903954	1.33E-04	0.0468674
TNFRSF10C	-0.62231	8.678819	-4.643669	6.54E-06	0.0122433
SLC46A2	-0.68594	5.88015	-4.085178	6.59E-05	0.0373188
LOC25489[illegible]	-0.71949	8.498379	-4.452628	1.48E-05	0.014977
ARHGEF40	-0.73174	6.880613	-3.995246	9.37E-05	0.0408941

Table 5: All 70 differentially expressed genes discovered with gcRMA noise correction in our analysis (JYo). 29 of the genes in this table are not included in the list that was uncovered using RMA-correction in Table 4

Functional Category	ID	Functional Term	Gene Count	adjusted p-value	Fold Change	Benjamini Score
BP	GO:0007067	mitotic nuclear division	8	1.8E-05	9.5031	0.010189
BP	GO:0000070	mitotic sister chromatid segregation	4	7.7E-05	47.135	0.022168
BP	GO:0051301	cell division	8	0.00016	6.7336	0.029505
BP	GO:0008283	cell proliferation	8	0.0002	6.4393	0.029505
CC	GO:0005876	spindle microtubule	4	0.00033	29.065	0.045679
BP	GO:0006955	immune response	8	0.00048	5.598	0.055034
CC	GO:0030496	midbody	5	0.00066	12.392	0.045679
BP	GO:0000086	G2/M transition of mitotic cell cycle	5	0.00112	10.752	0.10801
BP	GO:0007059	chromosome segregation	4	0.00151	17.329	0.124695
CC	GO:0005813	centrosome	7	0.00191	5.2536	0.066981
CC	GO:0000776	kinetochore	4	0.00198	15.789	0.066981
CC	GO:0000777	condensed chromosome kinetochore	4	0.00243	14.7	0.066981
CC	GO:0005654	nucleoplasm	18	0.00323	2.0672	0.074296
BP	GO:0000281	mitotic cytokinesis	3	0.00419	30.475	0.302441
CC	GO:0005819	spindle	4	0.00613	10.569	0.120782
CC	GO:0072686	mitotic spindle	3	0.00704	23.394	0.121486
BP	GO:0045740	positive regulation of DNA replication	3	0.00863	21.043	0.523985
BP	GO:0007077	mitotic nuclear envelope disassembly	3	0.00945	20.086	0.523985
MF	GO:0008009	chemokine activity	3	0.0115	18.132	1
CC	GO:0000228	nuclear chromosome	3	0.01156	18.097	0.173885
CC	GO:0005634	nucleus	26	0.0126	1.5351	0.173885
CC	GO:0005874	microtubule	5	0.01524	5.1402	0.191236
KEGG_PATHWAY	hsa05323	Rheumatoid arthritis	5	0.00047	13.028	0.044192
KEGG_PATHWAY	hsa04110	Cell cycle	5	0.00169	9.246	0.080047
INTERPRO	IPR001811	Chemokine interleukin-8-like domain	3	0.00821	21.614	0.709178
KEGG_PATHWAY	hsa04668	TNF signaling pathway	4	0.00997	8.572	0.261552
KEGG_PATHWAY	hsa04114	Oocyte meiosis	4	0.01101	8.2631	0.261552

Table 6: All significant functional terms analyzed using RMA-corrected data (JYe). More details will be covered in Jessica Yeh's paper.

Functional Category	ID	Functional Term	Gene Count	adjusted p-value	Fold Change	Benjamini Score
CC	GO:0005876	spindle microtubule	6	3E-07	41.418	3.16E-05
BP	GO:0051301	cell division	10	4E-06	7.8651	0.001968
BP	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	6	2E-05	17.953	0.004718
CC	GO:0000922	spindle pole	6	3E-05	16.719	0.001446
BP	GO:0007067	mitotic nuclear division	8	3E-05	8.88	0.004718
CC	GO:0030496	midbody	6	6E-05	14.127	0.002161
BP	GO:0030593	neutrophil chemotaxis	5	9E-05	20.854	0.008543
BP	GO:0000070	mitotic sister chromatid segregation	4	9E-05	44.045	0.008543
BP	GO:0007059	chromosome segregation	5	0.0001	20.241	0.008543
BP	GO:0070098	chemokine-mediated signaling pathway	5	0.0001	19.386	0.008667
CC	GO:0000776	kinetochore	5	0.0001	18.749	0.003661
BP	GO:0007080	mitotic metaphase plate congression	4	0.0003	29.76	0.019587
CC	GO:0072686	mitotic spindle	4	0.0003	29.633	0.006737
BP	GO:0007077	mitotic nuclear envelope disassembly	4	0.0005	25.025	0.028122
BP	GO:0055015	ventricular cardiac muscle cell development	3	0.0006	82.584	0.028122
MF	GO:0008009	chemokine activity	4	0.0007	22.967	0.089974
CC	GO:0005871	kinesin complex	4	0.0007	22.923	0.012002
BP	GO:0006955	immune response	8	0.0007	5.2309	0.033537
BP	GO:0051310	metaphase plate congression	3	0.0008	68.82	0.034213
BP	GO:0009612	response to mechanical stimulus	4	0.0012	18.663	0.047695
BP	GO:0000086	G2/M transition of mitotic cell cycle	5	0.0015	10.047	0.052571
BP	GO:0008283	cell proliferation	7	0.0019	5.2649	0.064615
MF	GO:0003682	chromatin binding	7	0.0024	5.037	0.123874
CC	GO:0005813	centrosome	7	0.0025	4.9909	0.038115
MF	GO:0003777	microtubule motor activity	4	0.0028	14.068	0.123874
BP	GO:0007018	microtubule-based movement	4	0.003	13.594	0.093747
BP	GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	4	0.0031	13.428	0.093747
CC	GO:0005874	microtubule	6	0.0033	5.8598	0.044198
BP	GO:0048247	lymphocyte chemotaxis	3	0.0045	29.494	0.125921
CC	GO:0005576	extracellular region	13	0.0049	2.4525	0.05778
BP	GO:0007062	sister chromatid cohesion	4	0.006	10.69	0.158867
MF	GO:0008017	microtubule binding	5	0.006	6.7632	0.169289
MF	GO:0004693	cyclin-dependent protein serine/threonine kinase activity	3	0.0063	24.825	0.169289
CC	GO:0005819	spindle	4	0.0071	10.041	0.075783
BP	GO:0002548	monocyte chemotaxis	3	0.0099	19.663	0.239098
BP	GO:0051726	regulation of cell cycle	4	0.0099	8.88	0.239098
BP	GO:0006954	inflammatory response	6	0.0113	4.358	0.248802
CC	GO:0000775	chromosome, centromeric region	3	0.0147	15.986	0.132609
CC	GO:0005829	cytosol	19	0.0149	1.7409	0.132609
BP	GO:0043547	positive regulation of GTPase activity	7	0.0152	3.4105	0.296691
BP	GO:0071346	cellular response to interferon-gamma	3	0.0177	14.488	0.331739
MF	GO:0008201	heparin binding	4	0.0184	7.0338	0.32498
MF	GO:0004197	cysteine-type endopeptidase activity	3	0.0193	13.837	0.32498
BP	GO:0060326	cell chemotaxis	3	0.0226	12.705	0.404778
BP	GO:0042493	response to drug	5	0.0232	4.5276	0.404778
BP	GO:0070374	positive regulation of ERK1 and ERK2 cascade	4	0.0247	6.2921	0.404778
BP	GO:0071347	cellular response to interleukin-1	3	0.0266	11.631	0.422184
CC	GO:0005654	nucleoplasm	16	0.029	1.7456	0.230654
INTERPRO	IPR019821	Kinesin, motor region, conserved site	4	0.0003	30.689	0.019184
INTERPRO	IPR001752	Kinesin, motor domain	4	0.0004	27.961	0.019184
INTERPRO	IPR001811	Chemokine interleukin-8-like domain	4	0.0004	27.353	0.019184
KEGG_PATH	hsa04114	Oocyte meiosis	5	0.0016	9.3898	0.107686
KEGG_PATH	hsa04110	Cell cycle	5	0.0024	8.4054	0.107686
INTERPRO	IPR000827	CC chemokine, conserved site	3	0.0025	39.32	0.091249
KEGG_PATH	hsa04060	Cytokine-cytokine receptor interaction	6	0.0049	5.147	0.142764
KEGG_PATH	hsa05323	Rheumatoid arthritis	4	0.0077	9.4752	0.164448
KEGG_PATH	hsa04062	Chemokine signaling pathway	5	0.0103	5.6036	0.164448
KEGG_PATH	hsa05142	Chagas disease (American trypanosomiasis)	4	0.0121	8.0175	0.164448
INTERPRO	IPR004827	Basic-leucine zipper domain	3	0.0124	17.476	0.297055
KEGG_PATH	hsa04668	TNF signaling pathway	4	0.0131	7.7927	0.164448

Table 7: All significant functional terms analyzed using gcRMA-corrected data. It is apparent that there are more significant terms in this table than Table 6 where RMA-correction is used.

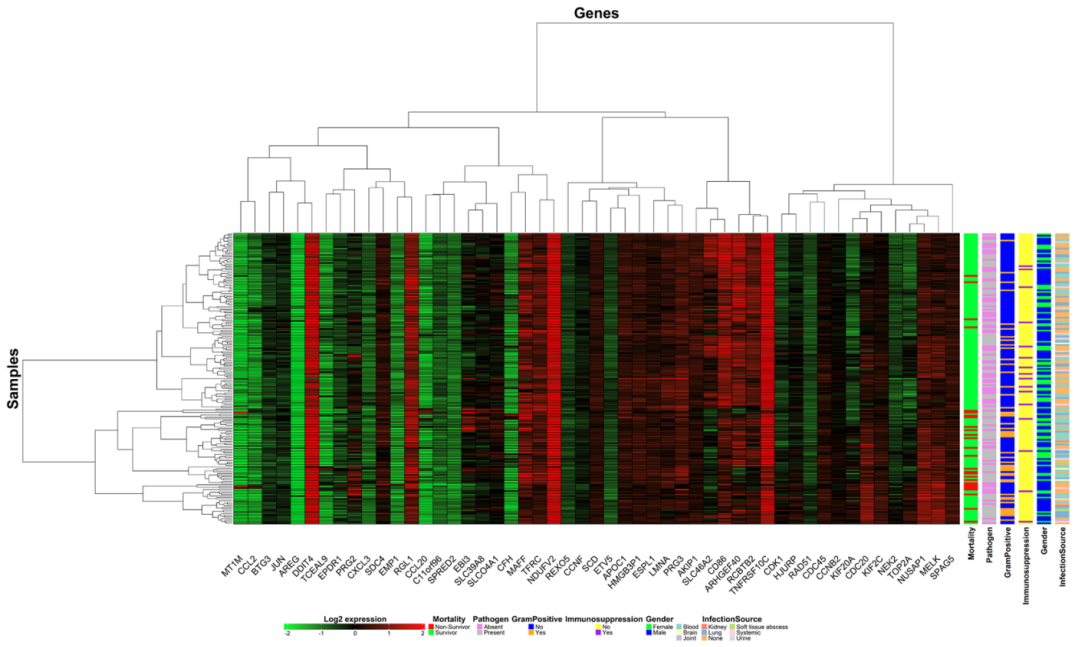

Figure 1: Heatmap of expression values produced in the original paper.

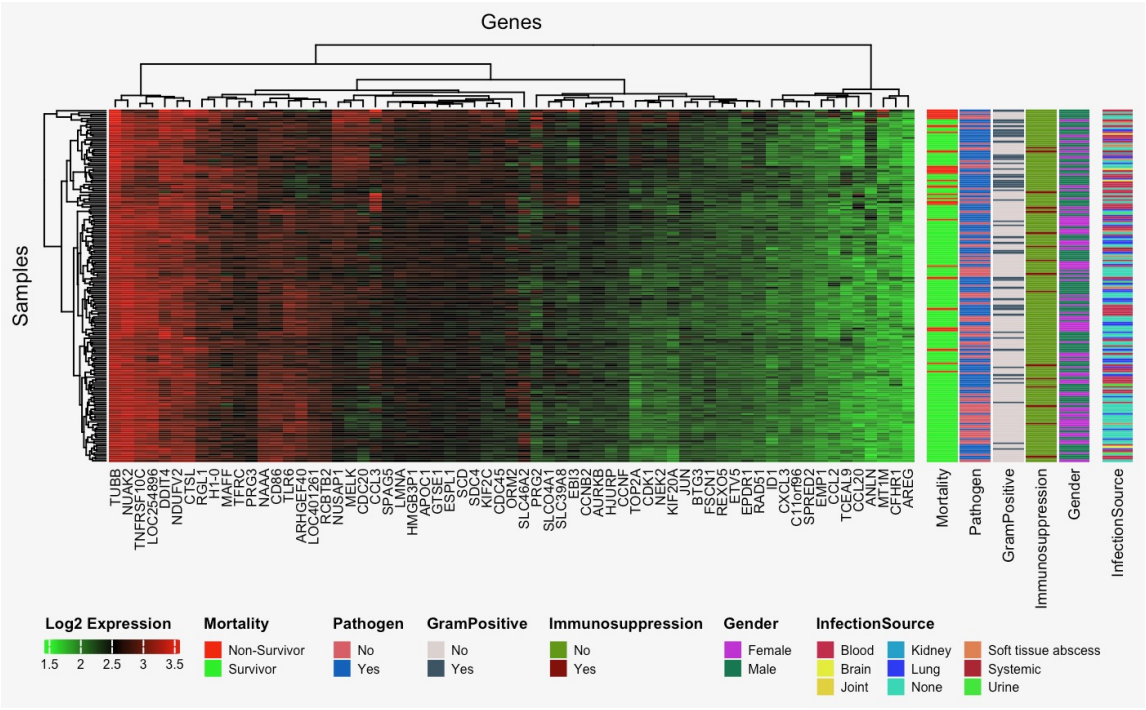

Figure 2: Our replicated heatmap using RMA (BZ). More details would be covered in Jessica Yeh's paper.

we extracted using RMA, with 29 genes from gcRMA-corrected data absent from the list of genes discovered using RMA-corrected data. We also find that gcRMA-processed data uncovered 70 genes, which is 5 more than RMA-processed data has uncovered. Similarly, we can also find through the functional enrichment analysis that gcRMA-corrected data generated more significant functional terms.

3.2 Potential Pitfalls

As shown in the previous section, different background noise correction methods of raw data can lead to very different results in terms of differentially expressed genes. Therefore, we cannot know whether or not our results are truly accurate until we generate a proper method to determine which noise correction method is optimal for this experiment. The authors seem to not have provided a rationale in the paper on why they chose to use RMA as their correction method.

Second, notice that in Table 5, some of the gcRMA-corrected differentially expressed genes are named as "Gene1|Gene2", with a "|" symbol in between the list of genes. This scenario occurs when we try to map the raw probe data to gene symbols, multiple gene symbols can get labeled to the same probe. Some possible solutions include randomly picking a gene from the multiple ones with same probe label. However, since the original study did not specify how we should handle this collison case, we just used "|" to to concatenate the list of genes into a label, which might have caused a slight shift in the reported statistics of our differential expression analysis.

3.3 Implications and Future Work

Our study have shown that using different background noise correction techniques when preprocessing raw data can produce different results for performing differential expression analysis and functional enrichment. Future studies need to take this difference into consideration when interpreting the results of differential expression analysis, and more research

can be performed to investigate why such a difference exists. Moreover, rationale on why a certain noise correction method is preferred over another should be provided to make the study fully convincing. For example, [6] suggested that two factors need to be taken into account when deciding on a background correction technique: (1) whether a study is aiming to detect differential gene expression between independent samples, (2) whether co-expressed genes should be identified.

4 Methods

4.1 Data Collection

We used the affymetrix microarray expression data collected from the NCBI Gene Expression Omnibus repository [7]. Our data contains gene expression profles of peripheral blood samples from 276 patients admitted to the PICU within the first 24 hours. Of the 276 patients, we filter out the control samples and the sample with only mild sepsis, leaving only 181 patients with septic shock, including 154 survivors and 27 non-survivors. The microarray data of each patient is represented by a file with .CEL extension. The GEO accession number for our dataset is GSE66099, and the data can be accessed through the ncbi website [8]. The data was collected from the Human Genome HG_U133_Plus_2 database [9]. For patient features such as gender, age, and pathogen involved in septic shock, we used the supplementary dataset, which is an excel file provided by the original study [2].

4.2 Background Correction and Normalization

The affy package [10, 11] was mostly used for loading the data and preprocessing it. As mentioned earlier, we experimented with the RMA [3] and gcRMA [4] background correction methods. We also use the Quantile Normalization method [12] to normalize our data.

4.3 Probe to Gene Mapping

We map the affymetrix probes to genes using the mappings provided by the Afymetrix Human Genome HG_U133_Plus_2 database, which can be accessed using the hgu133plus2.db package in R [9]. When a gene is mapped to different probes, the expression values from each probe is averaged to form the final gene expression value.

4.4 Differential Expression Analysis

We use the R Limma [13] package to perform differential expression analysis. Jae Yoon will discuss more about the methodology in his paper.

4.5 Functional Enrichment Analysis

We use the DAVID tool [14] to perform functional enrichment analysis. Jessica Yeh will discuss more about the methodology in her paper.

References

[1] Pediatric sepsis: Symptoms, diagnosis amp; treatment. Yale Medicine; 2019. Available from: https://www.yalemedicine.org/conditions/sepsis-in-kids.

[2] Mohammed A, Cui Y, Mas VR, Kamaleswaran R. Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients. Scientific Reports. 2019 Aug;9(1):11270. Available from: https://doi.org/10.1038/s41598-019-47703-6.

[3] Tai YC, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time course data. The Annals of Statistics. 2006;34(5):2387 2412. Available from: https://doi.org/10.1214/009053606000000759.

[4] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal

of the American Statistical Association. 2004;99(468):909-17. Available from: https://doi.org/10.1198/016214504000000683.

[5] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300. Available from: http://www.jstor.org/stable/2346101.

[6] Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics. 2007 07;23(13):i282-8. Available from: https://doi.org/10.1093/bioinformatics/btm201.

[7] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30 1:207-10.

[8] Wong H, Sweeney T. Unique Patients from the Genomics of Pediatric SIRS and Septic Shock Investigators (GPSSSI). U.S. National Library of Medicine; 2015. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66099.

[9] Carlson M. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2); 2016. R package version 3.2.3.

[10] Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15.

[11] MacDonald JW. affycoretools: Functions useful for those doing repetitive analyses with Affymetrix GeneChips; 2020. R package version 1.62.0.

[12] Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.

Bioinformatics (Oxford, England). 2003 January;19(2):185—193. Available from: https://doi.org/10.1093/bioinformatics/19.2.185.

[13] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.

[14] et al GD. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology. 2003.

Author Contributions

BZ obtained the raw data, performed data preprocessing, probe to gene mapping, and analyzed the effect of using RMA and gcRMA background noise collection. BZ also wrote code for Table 2 and created Figure 2. JYo performed differential gene expression analysis, helped visualize Table 2 in excel, and produced Tables 4 and 5. Jye performed functional enrichment analysis and produced Tables 6, 7.

Acknowledgements

We thank the course staff of QCB455 for providing guidance on our project and addressing our questions and confusions.