Name: Joshua Steven Sequeira

GUID: 3047039S

Course: BIOL5371 Statistics for Bioinformatics

Statistics and Data Analysis for Bioinformatics Assessment

Introduction

Arthritis refers to any disorder causing inflammation in the joints of patients. Gout is a common arthritis that occurs as a result of precipitation of monosodium urate in the joints due to hyperuricemia. This causes an autoimmune response leading to inflammation in the joints (Yu et al., 2023). Thus, the main genes involved in gout belong to the innate inflammatory pathways(Chang et al., 2022). Neutrophil extracellular traps are another important mechanism by which the immune system repairs gout flares (Lee et al., 2017).

Septic arthritis occurs when the joint inflammation is due to bacterial invasion (most commonly *Staphylococcus aureus*) of the synovium and joint space (Momodu & Savaliya, 2023). Neutrophils are instrumental in engulfing invading pathogens but are also the main cause of tissue destruction. Synovial fluid analysis and Gram staining are the most common methods of diagnosis(Boff et al., 2018).

Research Question: Gout and Septic Arthritis exhibit many clinical symptoms common to each other such as inflammation, fever, redness, and reduced movement in the diseased joints. Even gram staining and occurrence of monosodium urate crystals cannot be used to exclude infection due to a high false positive rate in these tests (Tzanis et al., 2022; Stirling et al., 2018). Due to the fast rate of spread and lethality of septic arthritis it is advisable to consider other methods of diagnosis such as analysis of RNAseq data.

Methods

All genes were identified by geneID to avoid problems due to duplicate gene symbols. In the top 50 most significant differentially expressed genes for gout there was only a single missing p-adjusted value (GCKR). No other missing values or duplicates impeded analysis in any way. All plots were carried out using ggplot2(*Wickham, 2016)* in R Statistical Software(*v4.4.1; R Core Team 2024*).

Parsing of given files

Files containing patient information and gene expression were merged by sampleID. Sex and Disease status were converted into factors. Both differential gene expression files were merged with each other by geneID. Genes were assumed to be significant if they had an absolute log2fold value > 1 and an adjusted p-value <0.05.

Analysis of Non-Gene Data

To identify the most significant differential indicators, the non-genetic factors were first taken. Sex and neutrophil count were analysed in three ways. Initially, their summary statistics were calculated across each sample disease group. Then, linear models were used to check if variation in neutrophil count could be explained by either the sex or disease status of the patients. To handle 0s an arbitrary value of 10^-100 was added to each data point. Finally, a boxplot was generated for neutrophil counts vs sex faceted by sample group.

Genetic Analysis

The genes were evaluated in four ways. First a hypergeometric test was performed to assess over-representation of differentially expressed genes in both groups. Then a principal component analysis was conducted to check for any clustering patterns. Finally, linear models were used to check if variation in gene expression for the top five most significant genes (in gout and septic arthritis separately) could be explained by the disease status, neutrophil data or sex of the patients. Their gene expression were also plotted as boxplots to compare the differences in expression.

Results

Parsing of Files

Based on the criteria above, the number of significant differentially expressed genes were as follows: 15 in Gout, 296 in Septic Arthritis.

The 8 that fulfilled the criteria for significance in both diseases were: MYO3B, SPP1, GATD3A, PCP2, KLHDC7A, EGFL6, RPL35P5, AC010300.1*(Refer Appendix).*

Analysis of Non-Gene Data

Summary Statistics: Each disease group had 9 samples and a sex ratio almost equal to 1. The mean neutrophil data were highest in the sepsis group (12.78) followed by gout (7.48) and lastly healthy controls (5.04). This seems to be biologically consistent because, as mentioned above, neutrophils play a role in both gout and septic arthritis but the autoimmune response is milder than the immune response to a bacterial infection (*Septic Arthritis vs. Gout*, 2023).

Linear Models: Both models were fitted to inverse gaussian distributions since neutrophil data consisted of positive, continuous covariates (gamma distributions fitted slightly worse).

Variation in Neutrophil data was not explained by sex ($F_{1,25}$= 0.0143, p=0.9056)
Variation in Neutrophil data was explained by Disease Group ($F_{2,24}$= 8.7824, p= 0.001373), echoing the conclusion reached from the summary statistics.

Boxplot:

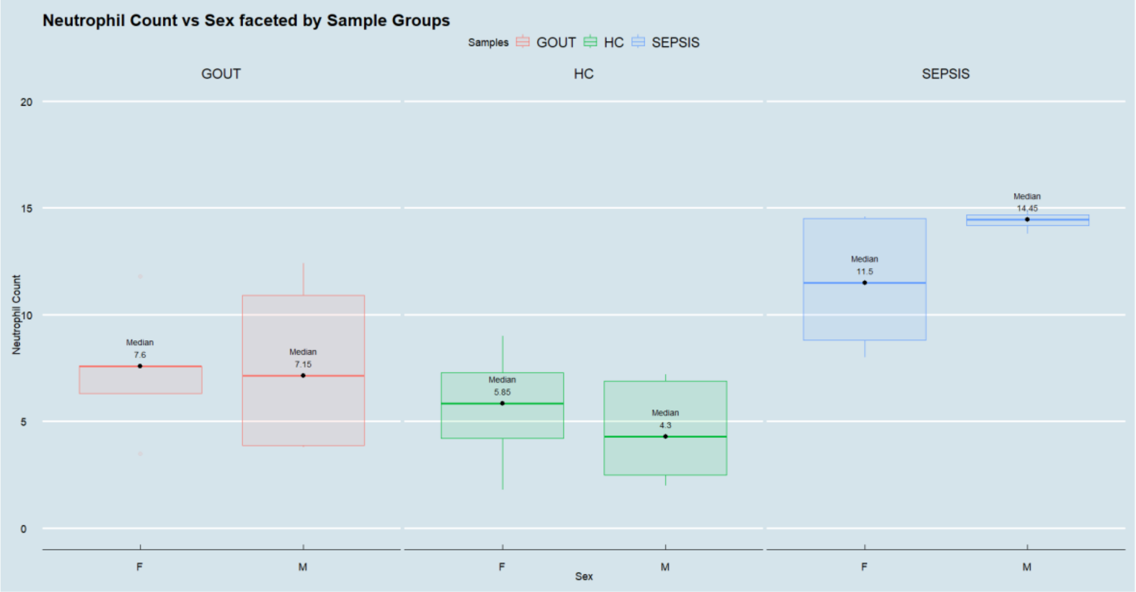

Fig 1: Boxplot of Neutrophil Count vs Sex Faceted by Disease Groups. Widest Interquartile ranges were seen in Male Gout and Female Sepsis Patients. Other conclusions are the same as above.

Genetic Analysis

Hypergeometric Test: Returned a p-value of 5.203284e-13. Thus, the shared genes are over-represented and there is common biology between gout and septic arthritis.

PCA: A plot of the first and second PCs covered 79.561% of the total variation. Slight overlap seen for gout and healthy samples. No significant patterns were seen in sepsis samples.

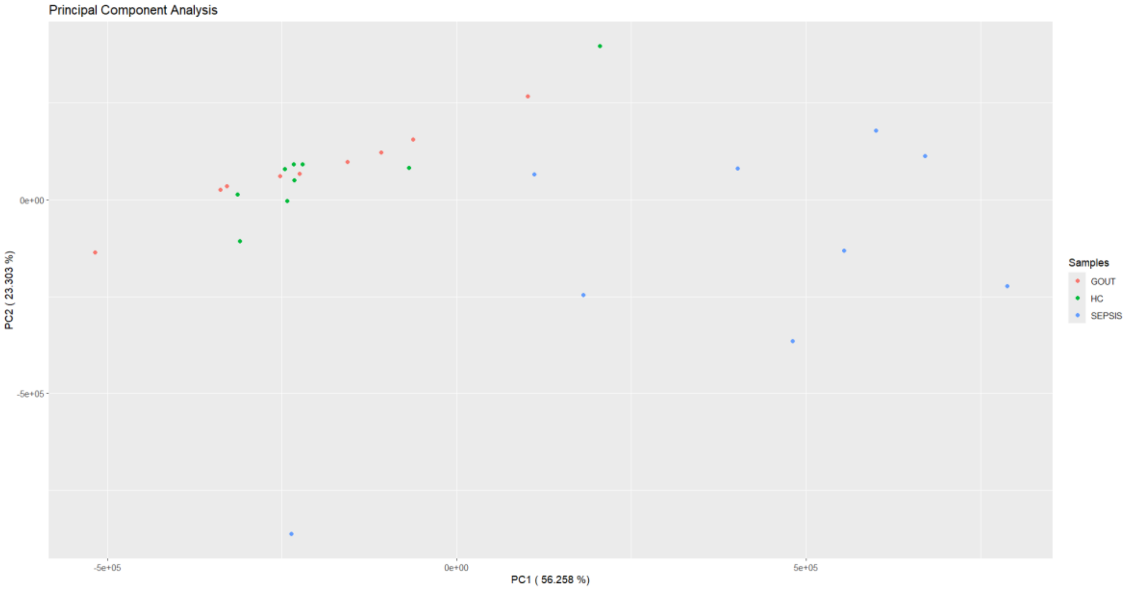

Fig 2: PCA for first two Principal Components colored by disease group

Linear Models:

All models were fitted to inverse gaussian distributions. In none of the cases were sex or neutrophil data found to significantly explain variation in gene expression. The models were then rerun with disease status as the only exploratory variable. The $F_{2,24}$ and p-values are given as follows

Table 1: Top 5 most significant differentially expressed genes in gout

Gene	**$F_{2,24}$**	**p-values**
SNORA73B	17.55702	2.005613e-05
KLHDC7A	19.63835	8.863563e-06
MYO3B	31.80688	1.785119e-07
EGFL6	10.57676	0.0005084353
SULT4A1	8.679736	0.001457589

Table 2: Top 5 most significant differentially expressed genes in SA

Gene	**$F_{2,24}$**	**p-values**
AKR1B10	94.82457	4.03758e-12
KYNU	179.0684	3.766245e-15
OASL	124.8675	2.063348e-13
TGM1	325.5045	4.081504e-18
RHCG	91.35173	6.002914e-12

Thus, the variation in the expression of these genes can be explained based on patient disease status.

Boxplots:

Boxplots were generated for both sets of 5 genes along with the 8 genes significant to both Gout and Septic Arthritis *(refer appendix)*. The main trend seen was that the five highly significant genes for septic arthritis were all upregulated with respect to both healthy and gout samples. The aim of this research is to identify genes which can differentiate healthy, gout and septic arthritis patients. However, most of the significant genes have similar expressions in two groups and an up/down regulated expression in the third.

Thus, I focus on only two genes SNORA73B and MYO3B. If we assume healthy control as a baseline expression, it is seen that one disease has upregulated levels of these genes while the other has downregulated levels. Thus, by quantifying the expression of these two genes, gout and septic arthritis can be differentiated both from healthy controls and from each other.

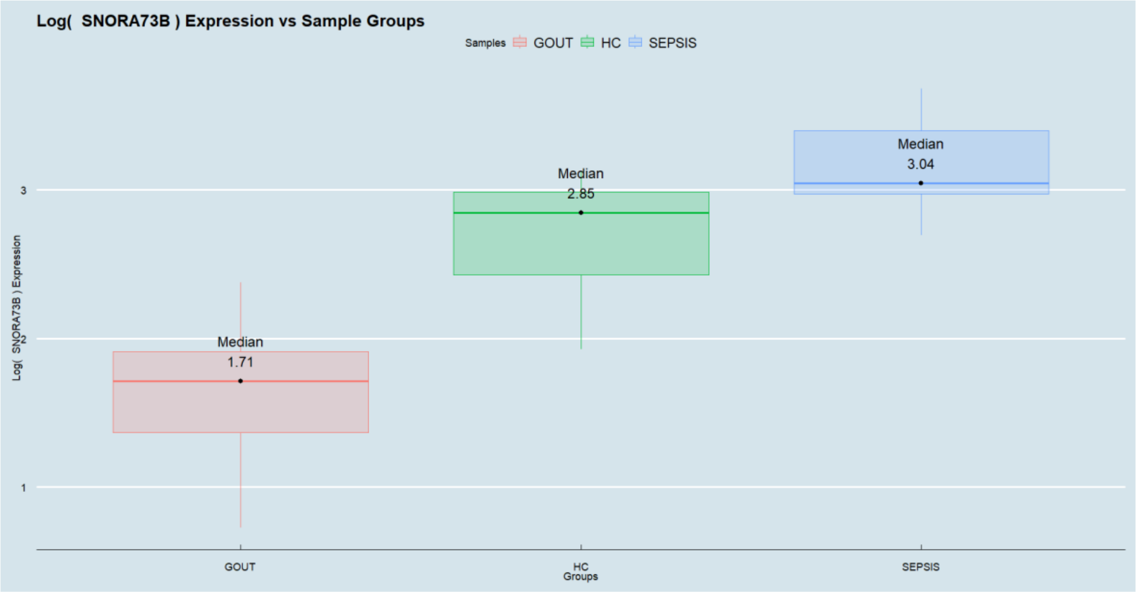

Fig 3: Boxplot of log of Gene Expression of SNORA73B vs Sample Groups

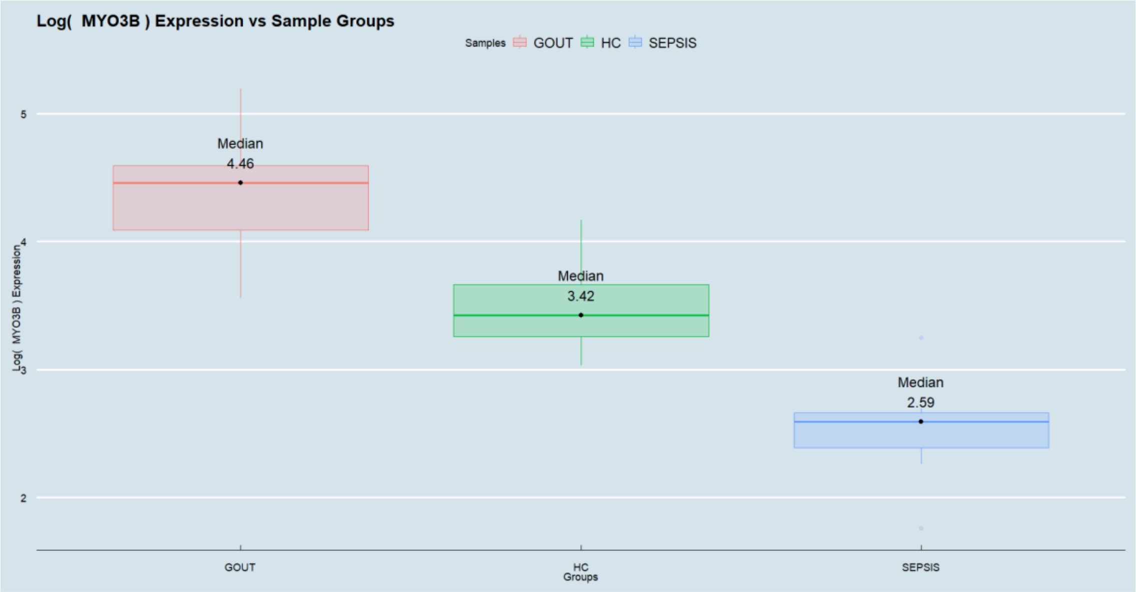

Fig 4: Boxplot of log of Gene Expression of MYO3B vs Sample Groups

Discussion

The clinical data given allows for multiple potential methods that may help to distinguish between gout and septic arthritis. Consistent with the linear models of neutrophil data vs disease status, a study shows that Delta neutrophil index (DNI) is related to burden of infection. Higher DNI is exhibited in septic arthritis as compared to gout patients and can aid in distinguishing them (Boff et al., 2018).

SNORA73 refers to small nucleolar RNA73 which belongs to the H/ACA class of snoRNAs and hosted in the small nucleolar RNA hosting gene 3 (SNHG3 was another gene significant to gout). A study shows that knockout of SNORA73 aids resistance of cells to oxidative stress, inflammation and cell death(Sletten et al., 2021) which is consistent with them being downregulated in gout. SNORAs were also found to be differentially expressed in older chondrocytes which are more predisposed towards septic arthritis(Balaskas et al., 2020)

MYO3B is a shorter isoform of class III myosin, which are actin-based motors with kinase domains in the amino terminus. They usually bind to cell-surface proteins. Differential expression of myosins has been linked to autoimmune diseases(Fu et al., 2023) which is consistent with it being upregulated in gout but not in septic arthritis.

However, it is important to note that each disease group in the dataset had only nine samples. Other important factors related to the various forms of arthritis such as age, are also missing from the dataset. These problems serves to diminish the significance of any conclusions made based on this dataset alone and thus further testing on larger datasets would be necessary to validate them.

References

Packages and Software

1. Ggfortify Package: Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
2. Ggplot2 Package: Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.
3. Ggthemes Package: Arnold J (2024). _ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'_. R package version 5.1.0, <https://CRAN.R-project.org/package=ggthemes>.
4. GridExtra Package: Auguie B (2017). _gridExtra: Miscellaneous Functions for "Grid" Graphics_. R package version 2.3, <https://CRAN.R-project.org/package=gridExtra>.
5. R statistical Software: R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Publications and Articles

1. Balaskas, P., Green, J. A., Haqqi, T. M., Dyer, P., Kharaz, Y. A., Fang, Y., Liu, X., Welting, T. J., & Peffers, M. J. (2020). Small Non-Coding RNAome of Ageing Chondrocytes. *International Journal of Molecular Sciences*, *21*(16), 5675. https://doi.org/10.3390/ijms21165675

2. Boff, D., Crijns, H., Teixeira, M. M., Amaral, F. A., & Proost, P. (2018). Neutrophils: Beneficial and Harmful Cells in Septic Arthritis. *International Journal of Molecular Sciences*, *19*(2), 468. https://doi.org/10.3390/ijms19020468

3. Chang, J.-G., Tu, S.-J., Huang, C.-M., Chen, Y.-C., Chiang, H.-S., Lee, Y.-T., Yen, J.-C., Lin, C.-L., Chung, C.-C., Liu, T.-C., & Chang, Y.-S. (2022). Single-cell RNA sequencing of immune cells in patients with acute gout. *Scientific Reports*, *12*, 22130. https://doi.org/10.1038/s41598-022-25871-2

4. Choi, U. Y., Kang, J.-S., Hwang, Y. S., & Kim, Y.-J. (2015). Oligoadenylate synthase-like (OASL) proteins: Dual functions and associations with diseases. *Experimental & Molecular Medicine*, *47*(3), e144. https://doi.org/10.1038/emm.2014.110

5. Fu, L., Zou, Y., Yu, B., Hong, D., Guan, T., Hu, J., Xu, Y., Wu, Y., Kou, J., & Lv, Y. (2023). Background and roles: Myosin in autoimmune diseases. *Frontiers in Cell and Developmental Biology*, *11*, 1220672. https://doi.org/10.3389/fcell.2023.1220672

6. Lee, K. H., Kronbichler, A., Park, D. D.-Y., Park, Y., Moon, H., Kim, H., Choi, J. H., Choi, Y., Shim, S., Lyu, I. S., Yun, B. H., Han, Y., Lee, D., Lee, S. Y., Yoo, B. H., Lee, K. H., Kim, T. L., Kim, H., Shim, J. S., … Shin, J. I. (2017). Neutrophil extracellular traps (NETs) in autoimmune diseases: A comprehensive review. *Autoimmunity Reviews*, *16*(11), 1160–1173. https://doi.org/10.1016/j.autrev.2017.09.012

7. Momodu, I. I., & Savaliya, V. (2023). Septic Arthritis. In *StatPearls [Internet]*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK538176/

8. *Septic Arthritis vs. Gout: Symptoms, Causes, and Treatment*. (2023, October 20). Healthline. https://www.healthline.com/health/septic-arthritis-vs-gout

9. Sletten, A. C., Davidson, J. W., Yagabasan, B., Moores, S., Schwaiger-Haber, M., Fujiwara, H., Gale, S., Jiang, X., Sidhu, R., Gelman, S. J., Zhao, S., Patti, G. J., Ory, D. S., & Schaffer, J. E. (2021). Loss of SNORA73 reprograms cellular metabolism and protects against steatohepatitis. *Nature Communications*, *12*, 5214. https://doi.org/10.1038/s41467-021-25457-y

10. Stirling, P., Tahir, M., & Atkinson, H. D. (2018). The Limitations of Gram-stain Microscopy of Synovial Fluid in Concomitant Septic and Crystal Arthritis. *Current Rheumatology Reviews*, *14*(3), 255–257. https://doi.org/10.2174/1573397113666170329123308

11. Tzanis, P., Klavdianou, K., Lazarini, A., Theotikos, E., Balanika, A., Fanouriakis, A., & Elezoglou, A. (2022). Septic Arthritis Complicating a Gout Flare: Report of Two Cases and Review of the Literature. *Mediterranean Journal of Rheumatology*, *33*(1), 75. https://doi.org/10.31138/mjr.33.1.75

12. Yu, H., Xue, W., Yu, H., Song, Y., Liu, X., Qin, L., Wang, S., Bao, H., Gu, H., Chen, G., Zhao, D., Tu, Y., Cheng, J., Wang, L., Ai, Z., Hu, D., Wang, L., & Peng, A. (2023). Single-cell transcriptomics reveals variations in monocytes and Tregs between gout flare and remission. *JCI Insight*, *8*(23), e171417. https://doi.org/10.1172/jci.insight.171417

Appendix

Model Diagnostics

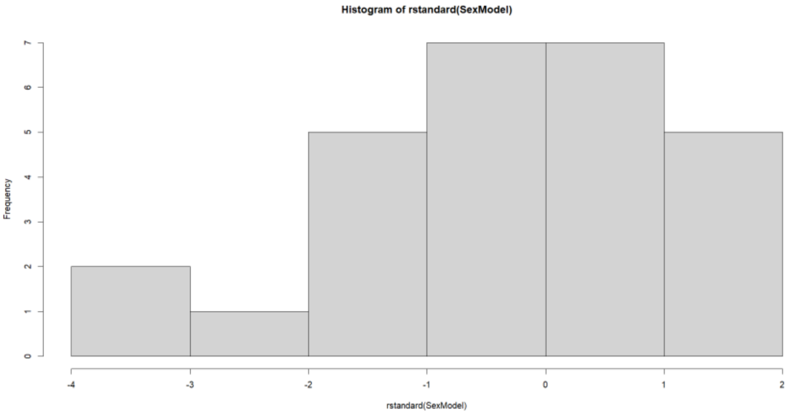

Fig 5: Neutrophils ~ Sex Model Residuals with Inverse Gaussian Fitting

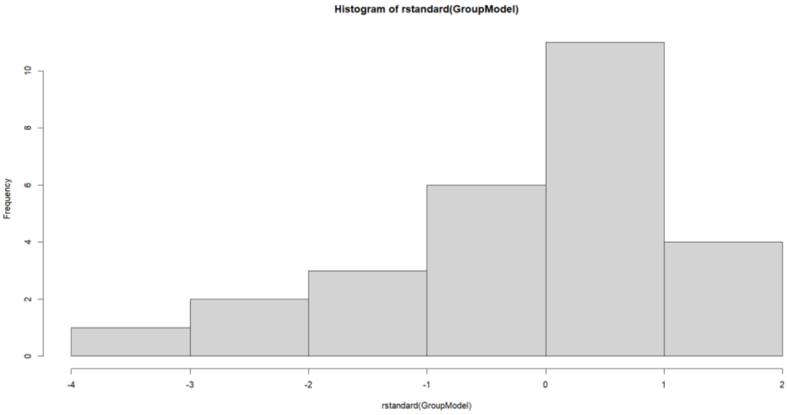

Fig 6: Neutrophils ~ Sample Group Model Residuals with Inverse Gaussian Fitting

Boxplots of the other Significant Genes Differentially Expressed in Both Disease Groups

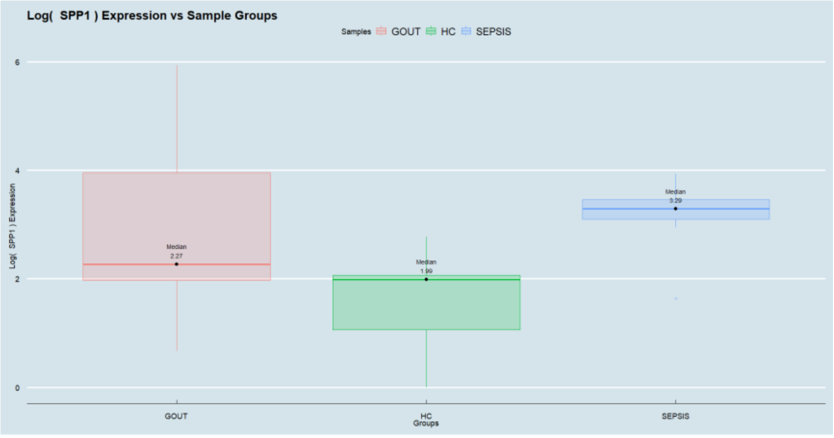

Fig 7: Boxplot of log of Gene Expression of SPP1 vs Sample Groups

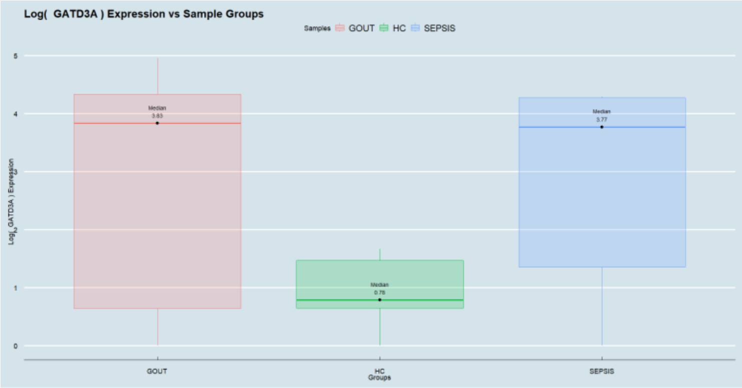

Fig 8: Boxplot of log of Gene Expression of GATD3A vs Sample Groups. Large Interquartile range in both diseases

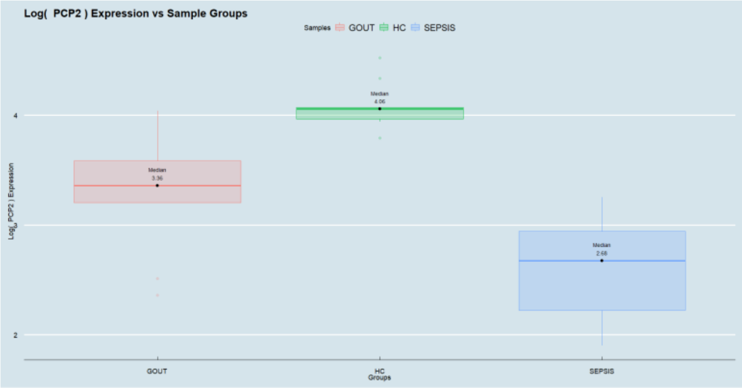

Fig 9: Boxplot of log of Gene Expression of PCP2 vs Sample Groups

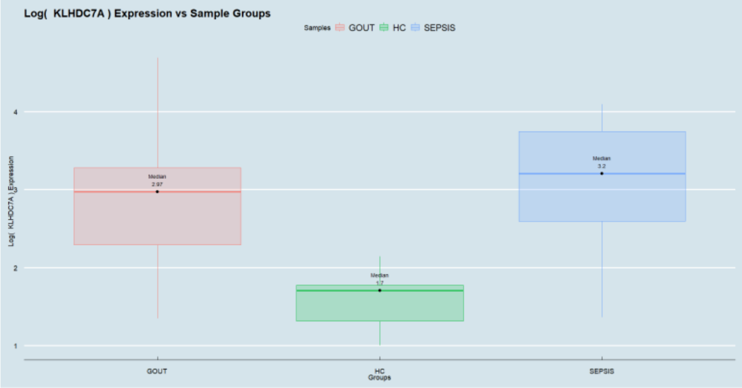

Fig 10: Boxplot of log of Gene Expression of KLHDC7A vs Sample Groups

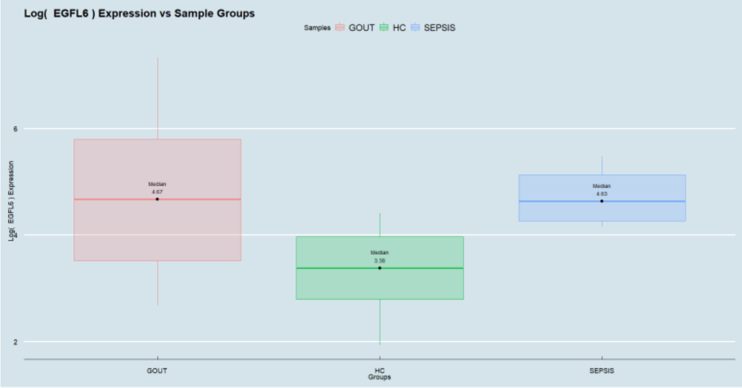

Fig 11: Boxplot of log of Gene Expression of EGFL6 vs Sample Groups

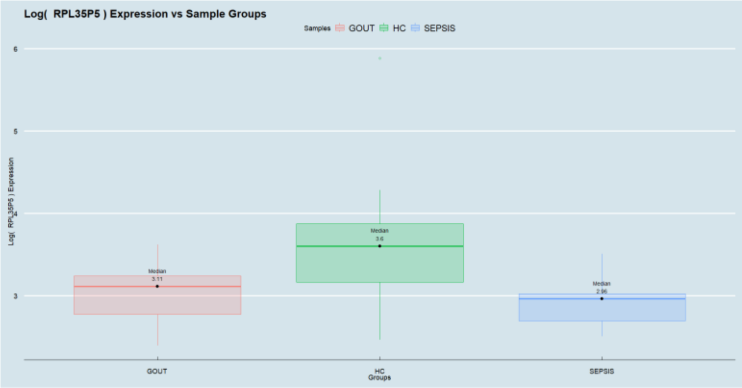

Fig 12: Boxplot of log of Gene Expression of RPL35P5 vs Sample Groups

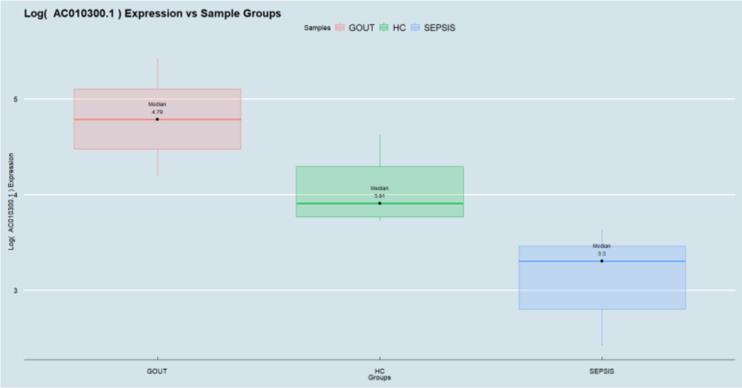

Fig 13: Boxplot of log of Gene Expression of AC010300.1 vs Sample Groups. This gene also presents a case where one disease has an upregulated value and the other has a downregulated expression. However, it has barely been studied and no biological relation could be found for its expression levels