Analysis of PBMC Gene Expression Data in Asthmatic Patients

Kieran Ford, Richard Garber and Jonathan Kahn

Contributing authors: kieran.ford@ufl.edu; richard.garber@ufl.edu; jonathan.kahn@ufl.edu;

Abstract

Asthma is a very common inflammatory disease affecting the lungs and reducing the quality of life of those who suffer it. Previous studies state that there is broad variability to how asthma presents itself throughout a sufferer's lifetime [1]. There are suggestions in previous research that asthma acquisition may be linked with altered immune system pathways [1–3], Previous studies have identified 73 differentially expressed genes compared between asthmatic and non-asthmatic patients [4]. However, these studies have not looked to identify a link specifically between the differentially expressed genes (DEGs) and the expression of cells of the immune system. In this study, we analyzed NGS data from peripheral blood mononuclear cells (PBMCs) among a group of 42 patients with asthma and 14 patients without asthma to identify any statistically significant differentially expressed genes (SSDEGs) and to conduct enrichment analysis and unsupervised clustering analysis to identify a link between the expression in the immune cells and asthma acquisition. Here we identify 65 DEGs but fail to find any SSDEGs. Enrichment analysis on the DEGs identified links to the biological processes of responses to fungus and other stimuli. This may suggest a link between the expression in the immune cells and asthma acquisition. Additionally, clustering analysis suggested no clear link between the expression of the immune cells and asthma acquisition. We expect our varying findings to act as a starting point for additional research into SSDEGs in immune cells relating to asthma acquisition and severity. Further experimentation with larger sample sizes will be required to further investigate links with a greater possibility of statistical significance.

1 Introduction

Asthma affects greater than 300 million individuals which is predicted to increase as more countries around the world adopt lifestyles similar to developed western countries [1]. There are many suggestions about what causes asthma to develop in humans and at what stage. Some previous studies have identified certain immune responses to fungus as having an association with asthma severity as well as a possible relationship to inflammation in pathways with helper T cells [3, 5]. Other genome-wide association studies (GWAS) also identified differentially expressed genes in asthma patients and loci of interest in the immune system [4, 6]. In this study we focused on supporting these suggestions and findings of possible links between immune cells and asthma acquisition by focusing on a more simplified level of analysis. The research question we sought to answer was, What are the differences in gene expression within peripheral blood mononuclear cells between patients with moderate or severe asthma and healthy patients? To adequately answer this question, we used a next generation sequencing (NGS) counts dataset from the NCBI GEO database, GSE207751, which contained a total of 56 samples – 42 asthma patients and 14 patients without asthma. The samples came from PBMCs and the asthma patients had varying levels of severity identified in their metadata. The metadata also included other characteristics of the patients such as age and gender. The data was collected from peripheral blood mononuclear cells (PBMCs) which are cells with a round nucleus in the bloodstream, with most being lymphocytes and monocytes [7]. These cells make up a large part of the immune system and T cells are a type of lymphocyte. Altered immune systems could be displayed via differentially expressed genes within this class of cells, which we can identify through our counts data and run through enrichment analysis to identify the pathways of these differentially expressed genes.

2 Methods

All code can be found and the GitHub repository here: https://github.com/rgarber11/GeGnomes-BioInformatics-Project
Before analysis could be done using the data set, there was some formatting that needed to be done. The data set identified genes by Ensembl ID, these needed to be replaced with HGNC symbols, this was done using the `biomaRt` library in R [8]. The counts were then TPM normalized, and saved as fixed_normalized_counts.csv. A log-scaled and TPM normalized version was also created.
`DESeq2` was used to apply a VST and normalize data. This VST-normalized data was used to generate both the PCA plot and the assay data for M3C's TSNE plot [9–11]. All plots were also created using `ggplot2`. For differential expression a volcano plot was created using DESeq data and a pCutoff value of 0.01. A heatmap [12] was created using the log scaled counts data of genes that were found to be differentially expressed, with an annotation dividing samples into healthy and asthmatic groups. Enrichment analysis on all genes

with a logFoldChange greater than 1 is performed with the GO biological processes, the GO cellular components, and the REAC databases using the g:Profiler tool [13–15].
Clustering was performed three times on the dataset using three different methods: PAM clustering, K-means [16], and `ConsensusClusterPlus` [17]. To perform clustering an additional column was added to the counts data that contains variance and counts were ordered by variance. Finally the data was broken into smaller sets so clustering could be done with different amounts of genes. For each clustering method, different k values were tested, in the end PAM and K-means clustering was done with a k value of 2. Consensus clustering was done with a max K value of 20, the hierarchical clustering method, the Pearson method of determining distance, 1000 repetitions, and with 80% of items sampled and 100% of features sampled. Sankey diagrams were created for each clustering method comparing the results of clustering with only 10, 100, 1000, 5000, and 10000 genes using the `ggalluvial` library [18]. The results of these clustering methods were used to annotate a heatmap [12] of the top five thousand genes by variance, showing which samples The results of clustering were compared against each other and the original condition using a chi square test. They were then adjusted due to there being 6 different tests, and interpreted with a cutoff of $p > 0.05$.

3 Results

Overall, we were unable to find a statistically significant difference in gene expression between people with or without asthma. While the dataset contained some variability as seen in Fig. 1, a PCA plot and t-SNE show that a majority of the variance does not come from whether participants had asthma, as seen in Figs. 2 and 3. Meanwhile, as can be seen in Table 1, DeSeq analysis shows that while certain singular tests would be independently significant, no test meets the barrier of $p < 0.5$ for statistical significance when taken together. That none of the genes expressed are significant can also be seen in Figs. 4 and 5. Thus, while certain genes are differentially expressed, none meet the criterion to give an affirmative answer to the experimental question.

Gene	baseMean	log2FoldChange	lfcSE	pvalue	padj	threshold
LRRC58-DT	1.0412624	4.321867	1.4174352	4.371686e-02	0.4221497	FALSE
MTND1P23	1.6782038	4.147562	1.2979116	3.610390e-02	0.3978533	FALSE
ADAMTS2	13.4136189	3.982281	0.9256661	5.165539e-05	0.2052859	FALSE
RPS20P13	0.6143606	3.710406	2.1985731	3.435173e-03	0.2504310	FALSE
SLC6A19	16.7556055	3.536763	0.3928395	9.961202e-01	0.9994219	FALSE
MTCO1P12	85.5174753	3.021756	0.5250292	8.666273e-03	0.2780836	FALSE

Table 1 DeSeq data showing statistical significance of the most differentially expressed genes.

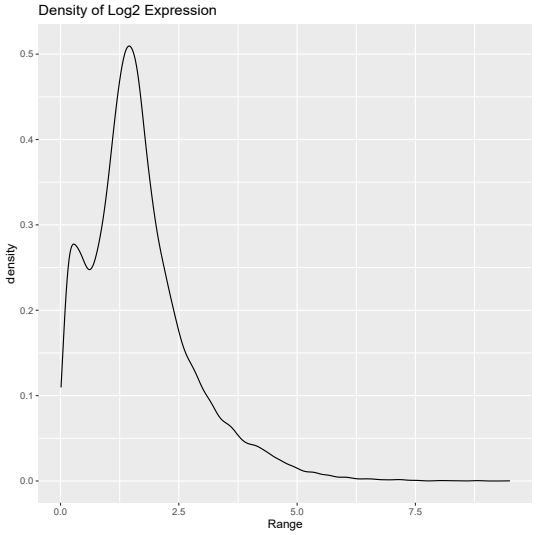


Fig. 1 Log-Range Density of Gene Expression

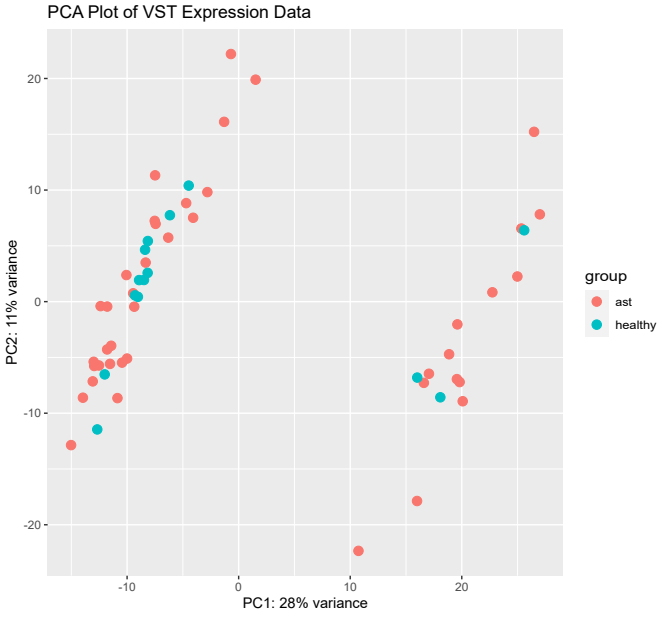


Fig. 2 PCA Plot of DeSeq of Gene Expression Data

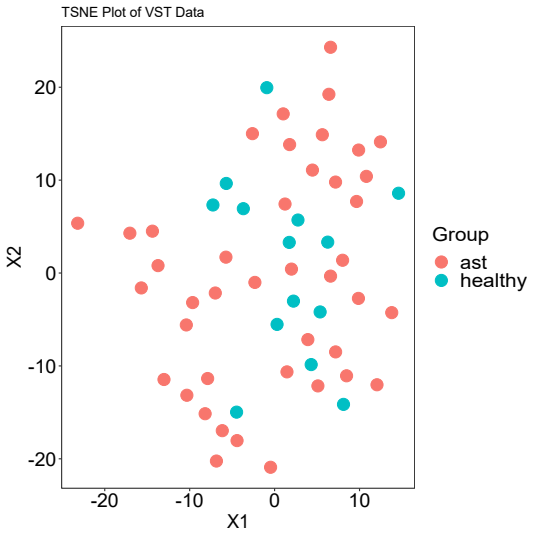


Fig. 3 t-SNE Plot of DeSeq Data

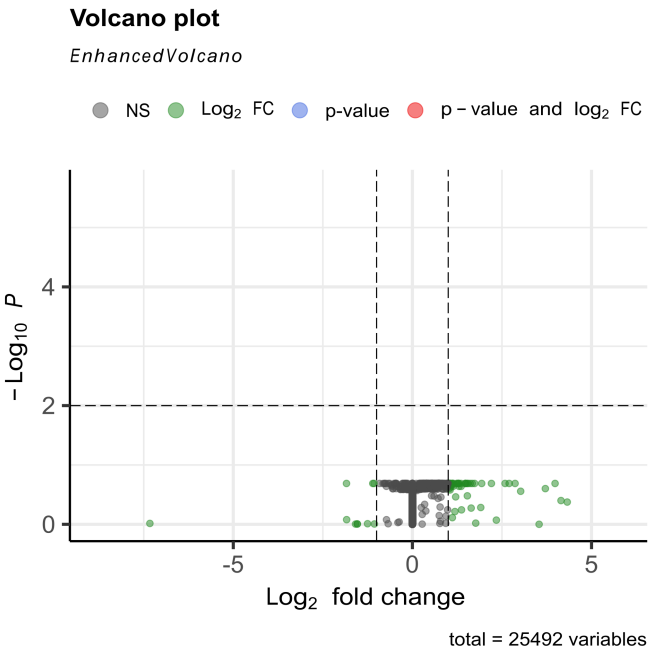


Fig. 4 Volcano Plot showing statistical significance and log-Fold Change.

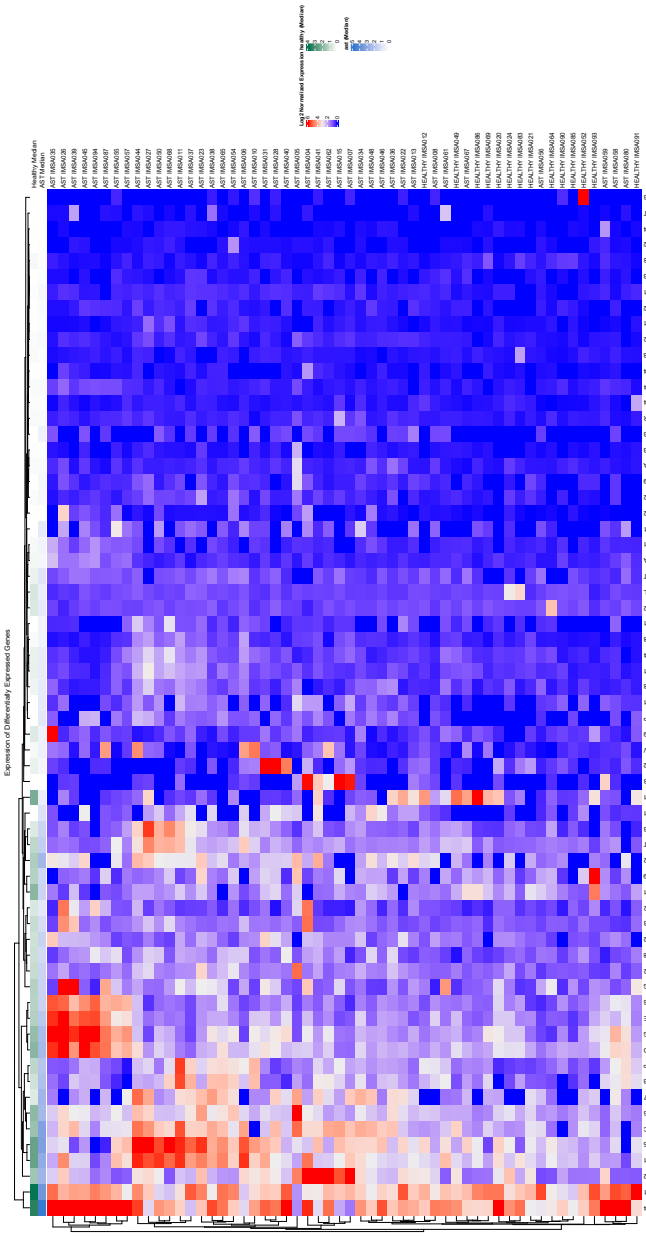


Fig. 5 Heatmap showing the expression of genes with highest logFold Change

Cluster analysis corroborates the results from other tests regarding statistical significance, while showing that analyzing different amounts of genes leads to varying effectiveness of confounding variables. Neither `ConsensusClusterPlus` with Hierarchical clustering, PAM, nor KMeans show significant statistical dependence with the experimental condition, while showing statistical dependence among each other. (Table 2) Beyond this, while two methods had the same optimal number of clusters as the condition (Fig. 6), Fig. 7 shows that ConsensusClusterPlus had significantly more, at 10 clusters. In addition, all methods show differences in sample clustering depending on the amount of genes used for the cluster analysis (Figs. 8, 10 and 9), which suggest that questions regarding the expression of specific sets of genes may result in varied conclusions, with different confounding variables.

	ConditionPAM	ConditionKMeans	ConditionCCP	PAMKMeans	PAMCCP	KMeansCCP
p-value	0.03250944	0.6325851	0.2205631	1.448363e-07	1.593233e-06	3.420940e-06
p-adj	0.09752833	0.6325851	0.4411262	8.690175e-07	7.966164e-06	1.368376e-05

Table 2 Statistical Independence of Clustering Algorithms when compared amongst themselves and the experimental condition.

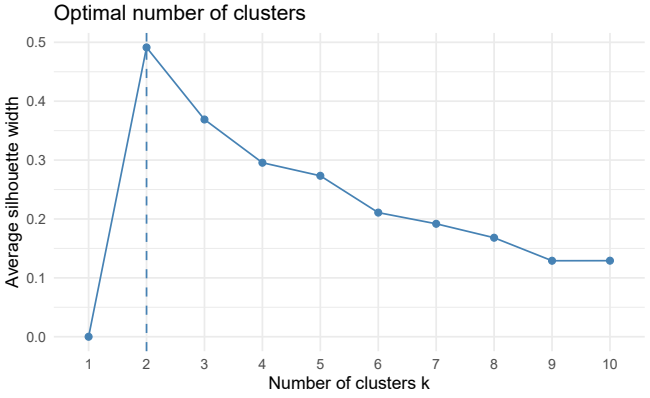


Fig. 6 Optimal Amount of Clusters as determined by `factoextra` [19]

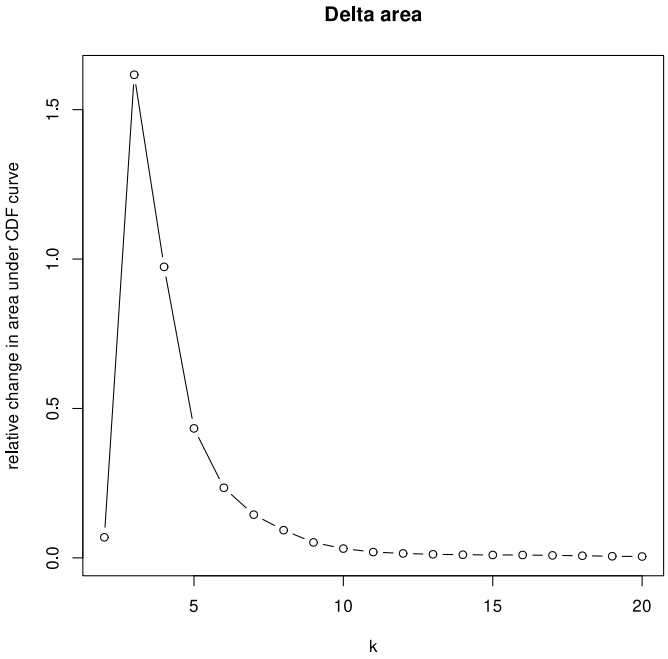


Fig. 7 Additional variability accounted for by additional clusters

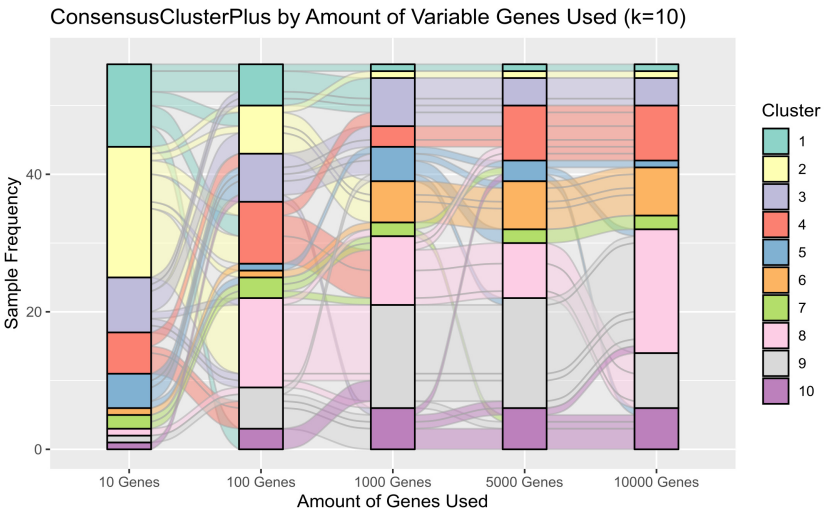


Fig. 8 Differences in `ConsensusClusterPlus` sample clusters when more genes are used in cluster analysis.

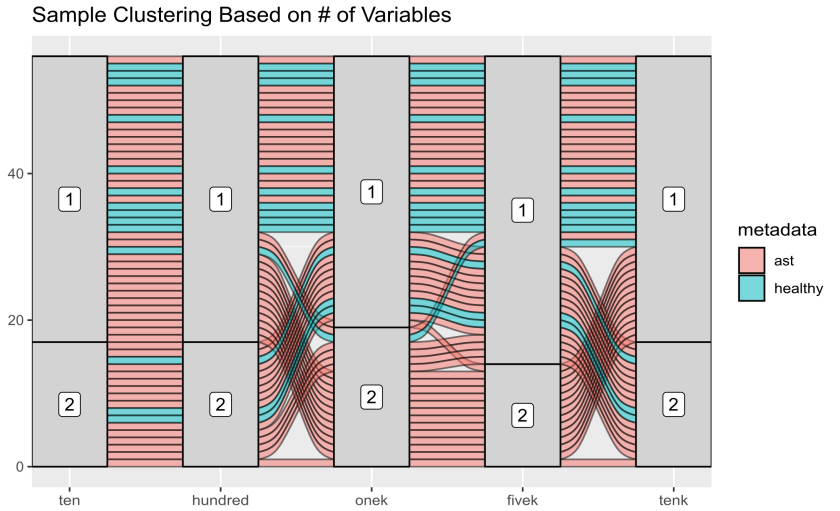


Fig. 9 Differences in PAM sample clusters when more genes are used in cluster analysis.

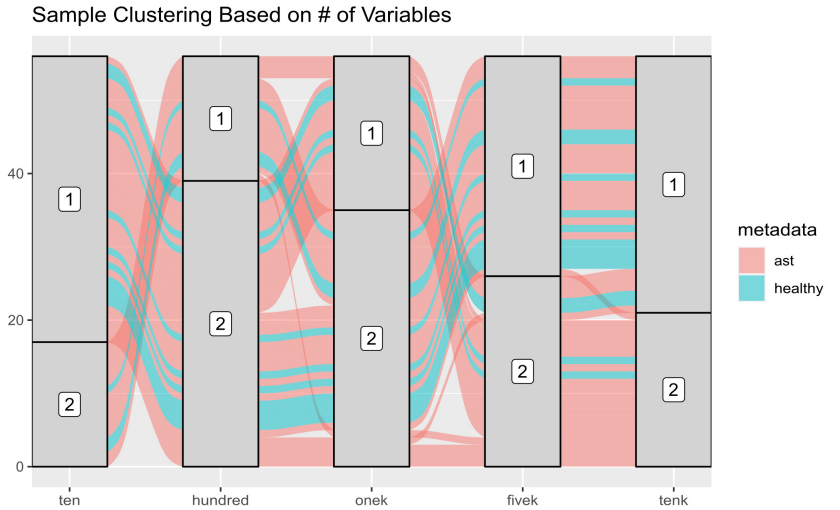
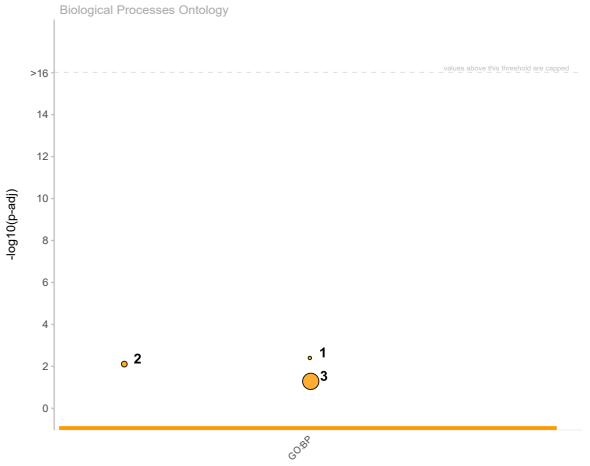


Fig. 10 Differences in KMeans sample clusters when more genes are used in cluster analysis.

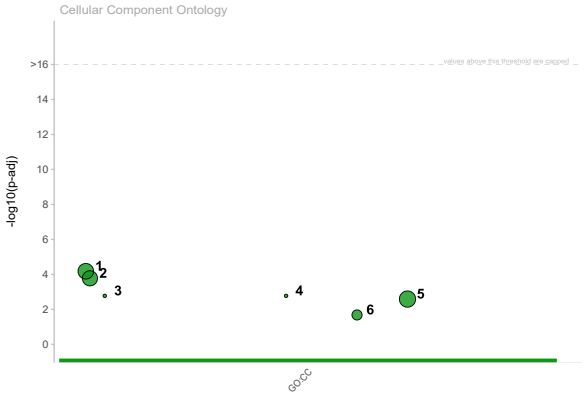
Finally, while independent analysis of the dataset reveals no significant differences in expression, enrichment suggests certain processes and components with which the gene expressions found within the dataset correlate. Enriching with the GO:Biological Processes database [14] (Fig. 11) included biological processes relating to different responses to fungus and stimulus in general, while the GO:Cellular Components and REACTOME database [14, 15] (Figs. 12 and 13) show correlation with cellular components involved in the lysosome, azurophilic granules, other antimicrobial peptides and extracellular space. While some of these tasks are the primary purposes of PBMCs, their p-values suggest that these pathways could be an important area of further research. However, analysis of this dataset did not find a statistically significant difference in gene expression between patients with or without asthma.



id	source	term_id	term_name	term_size	p_value
1	GO:BP	GO:0050832	defense response to fungus	55	3.8e-03
2	GO:BP	GO:0009620	response to fungus	65	7.4e-03
3	GO:BP	GO:0050896	response to stimulus	8885	4.9e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

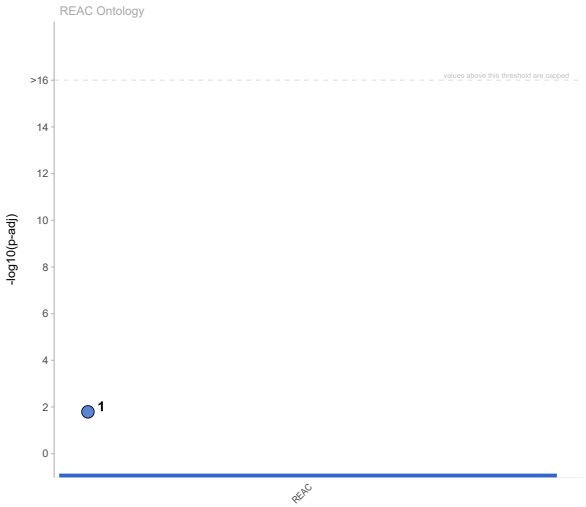
Fig. 11 `g:Profiler` enrichment analysis of differentially expressed genes using the GO Biological Processes Database



id	source	term_id	term_name	term_size	p_value
1	GO:CC	GO:0005576	extracellular region	4315	6.3e-05
2	GO:CC	GO:0005615	extracellular space	3364	1.6e-04
3	GO:CC	GO:0005766	primary lysosome	154	1.6e-03
4	GO:CC	GO:0042582	azurophil granule	154	1.6e-03
5	GO:CC	GO:0071944	cell periphery	6237	2.4e-03
6	GO:CC	GO:0062023	collagen-containing extracellular matrix	423	2.0e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

Fig. 12 `g:Profiler` enrichment analysis of differentially expressed genes using the GO Cellular Components Database



id	source	term_id	term_name	term_size	p_value
1	REAC	REAC:R-HSA-6803157	Antimicrobial peptides	95	1.6e-02

g:Profiler (biit.cs.ut.ee/gprofiler)

Fig. 13 `g:Profiler` enrichment analysis of differentially expressed genes using the REACTOME Database

4 Conclusion

Overall, statistical analysis did not find a significant difference in gene expression between asthmatic and healthy patients. There were no genes whose difference in expression were statistically significant per DeSeq analysis. Clustering analysis also showed that much of the variance in gene expression is not directly related to the experimental condition. The likely reason that the analysis of the data did not provide meaningful results was the limited sample size. While there were many genes in the data set, samples were taken from only 56 different patients, leading to low statistical power. Analysis of gene expression of a larger sample size will yield more meaningful results. Other problems with our study include the difficulty in studying a disease as variable as asthma. As such, we expect a high level of variability within our experimental group, making analysis difficult. Finally, our dataset may suffer from selection biases, as it comes from a single lab in Massachusetts. As such, results may not be applicable to other parts of the world, especially with asthma often being related to environmental factors.

However, the analysis still showed interesting paths for further investigation. Enrichment analysis identified genes that correspond to biological processes involving response to stimulus, and response to fungi. These biological processes and cellular components have previously been correlated with asthma [3, 5], and this analysis may point to specific genes and pathways which cause these correlations to occur. These results from enrichment analysis suggest the need for further investigation, which may uncover specific pathways that lead to an asthmatic response to immune stimuli and fungi. However with the current data set, no significant difference or relationship between the gene expression of healthy and asthmatic patients' PBMC cells could be found. Due to the limitations on our dataset, we would need to generate additional experimentation on a larger sample of patients in order to possibly identify SSDEGs and further investigate any link between the expression of the immune cells and asthma acquisition in a setting where statistical significance can be supported.

References

[1] Holgate, S.T., Wenzel, S., Postma, D.S., Weiss, S.T., Renz, H., Sly, P.D.: Asthma. Nat. Rev. Dis. Primers **1**, 15025 (2015)

[2] Camiolo, M.J., Kale, S.L., Oriss, T.B., Gauthier, M., Ray, A.: Immune responses and exacerbations in severe asthma. Curr. Opin. Immunol. **72**, 34–42 (2021)

[3] Holgate, S.T.: Innate and adaptive immune responses in asthma. Nat. Med. **18**(5), 673–683 (2012)

[4] Cao, X., Ding, L., Mersha, T.B.: Development and validation of an RNA-seq-based transcriptomic risk score for asthma. Sci. Rep. **12**(1), 8643 (2022)

[5] Denning, D.W., O'Driscoll, B.R., Hogaboam, C.M., Bowyer, P., Niven, R.M.: The link between fungi and severe asthma: a summary of the evidence. Eur. Respir. J. **27**(3), 615–626 (2006)

[6] Zhu, Z., Lee, P.H., Chaffin, M.D., Chung, W., Loh, P.-R., Lu, Q., Christiani, D.C., Liang, L.: A genome-wide cross-trait analysis from UK biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat. Genet. **50**(6), 857–864 (2018)

[7] Pourahmad, J., Salimi, A.: Isolated human peripheral blood mononuclear cell (PBMC), a cost effective tool for predicting immunosuppressive effects of drugs and xenobiotics. Iran. J. Pharm. Res. **14**(4), 979 (2015)

[8] Steffen Durinck ¡biomartdev@gmail.com¿, Wolfgang Huber: biomaRt. Bioconductor (2017)

[9] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014)

[10] Zhu, A., Ibrahim, J.G., Love, M.I.: Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics **35**(12), 2084–2092 (2019)

[11] John, C.R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., Barnes, M.: M3C: Monte Carlo reference-based consensus clustering (2018)

[12] Gu, Z., Eils, R., Schlesner, M.: Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics **32**(18), 2847–2849 (2016)

[13] Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., Vilo, J.: g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. **47**(W1), 191–198 (2019)

[14] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[15] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., *et al.*: The reactome pathway knowledgebase. Nucleic acids research **46**(D1), 649–655 (2018)

[16] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: Cluster: Cluster Analysis Basics and Extensions. (2022). R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source). https://CRAN.R-project.org/package=cluster

[17] Wilkerson, M.D., Hayes, D.N.: ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics **26**(12), 1572–1573 (2010)

[18] Brunson, J.C.: ggalluvial: Layered grammar for alluvial plots. Journal of Open Source Software **5**(49), 2017 (2020). https://doi.org/10.21105/joss.02017

[19] Kassambara, A., Mundt, F.: Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2020). R package version 1.0.7. https://CRAN.R-project.org/package=factoextra