

Identifying Asthma genetic signature patterns by mining Gene Expression *BIG* Datasets using Image Filtering Algorithms

Mahmood Yaseen Hachim¹, Bassam Mahboub^{1,2} MD, Qutayba Hamid¹ PhD, MD, FRS, Rifat Hamoudi^{1,3*}, PhD, CEng, MIEEE, MBCS

¹ Sharjah Institute for Medical Research, College of Medicine, University of Sharjah, United Arab Emirates, 2 Rashid Hospital, Dubai, United Arab Emirates 3 Division of Surgery and Interventional Science, University College London, United Kingdom.

Abstract— Asthma is a treatable but incurable chronic inflammatory disease affecting more than 14% of the UAE population. Asthma is still a clinical dilemma as there is no proper clinical definition of asthma, unknown definitive underlying mechanisms, no objective prognostic tool nor bedside noninvasive diagnostic test to predict complication or exacerbation. Big Data in the form of publicly available transcriptomics can be a valuable source to decipher complex diseases like asthma. Such an approach is hindered by technical variations between different studies that may mask the real biological variations and meaningful, robust findings. A large number of datasets of gene expression microarray images need a powerful tool to properly translate the image intensities into truly differential expressed genes between conditioned examined from the noise. Here we used a novel bioinformatic method based on the coefficient of variance to filter nonvariant probes with stringent image analysis processing between asthmatic and healthy to increase the power of identifying accurate signals hidden within the heterogeneous nature of asthma. Our analysis identified important signaling pathways members, namely NFKB and TGFB pathways, to be differentially expressed between severe asthma and healthy controls. Those vital pathways represent potential targets for future asthma treatment and can serve as reliable biomarkers for asthma severity. Proper image analysis for the publicly available microarray transcriptomics data increased its usefulness to decipher asthma and identify genuine differentially expressed genes that can be validated across different datasets.

Keywords—microarray, image analysis, transcriptomics, normalization, coefficient of variance

I. INTRODUCTION

Asthma is one of the most common chronic inflammatory disease, usually characterized by being a lifelong condition with high disease burden [1]. Bronchial hyperresponsiveness, inflammation, and airway obstruction episodes are the main characteristic features of this disease [2]. The prevalence of asthma was shown to be increased in the last years, reaching alarming levels [3]. Many theories on the factors that made people at risk of developing asthma were suggested; however, most of these theories had no conclusive results [3].

Dr Rifat Hamoudi is funded by the Sharjah Research Academy (Grant code: MED001) and Al-Jalila Foundation (Grant code: AJF201741).

* Corresponding and Senior Author

Many reasons might play a role in hindering the reach for a proper understanding of what makes people develop asthma, or what convert entirely controlled asthmatic cases into severe or fatal ones. [4]. The fact that different regions of the asthmatic lung can react differently to the same provoking factor makes the patients clinical features and response to therapy differ giving rise to the asthma heterogeneity.

The use of Omics technologies, including genomics, transcriptomics, proteomics, and metabolomics, provided extraordinarily details and enriched our knowledge about the asthma heterogeneity and molecular basis of asthma [5]. Moreover, these techniques helped in the discovery of new markers essential for precise classification of asthma as well as the development of new therapeutic targets [6]. However, the use of these techniques gives no conclusive and even contradictory results [7]. The entire biological complexity of asthma cannot be captured using an isolated output from these sophisticated technologies if taken separately [8].

Among Omics, transcriptomic profiling of bronchial biopsies as well as epithelial brushing cells showed the potential to discover gene expression profiles that are characteristic in asthma [9] and can identify different molecular mechanisms that separate asthmatic phenotypes [10]. Moreover, this approach is believed to have the power to identify novel clinical biomarkers to help physicians in making more precise therapeutic strategies in treating individual patients [11]. However, such an approach in a broader range of patients has not yet been performed [12]. This costly approach can be more informative if a large number of samples are combined to extract meaningful information by using a large number of datasets available in public databases [13]. Such motivation is faced by the fact that microarrays studies showed an extreme method to method variations in the results of the significant differentially expressed genes (DEG). The main reason behind this is the variability in experimental processes that can mask the investigated biological effect [14].

Most of the huge publicly available transcriptomics datasets were made by arrays that can detect thousands of genes simultaneously. Microarray technology presents the expression level of a gene as pixel intensity. Translating pixel intensities

into transcript expression requires a series of computations [15], and each of these steps carries its noise and errors. The scanned signal consists of true spot intensity and noise referred to as background, so even a small error in the estimate of true or noise signal can influence the results and shift an upregulated gene to be referred as downregulated genes or not affected by the condition examined[16].

Ideally, identifying truly DEG among thousands of genes needs to reduce the number of genes that are equally expressed between disease and control by a filtration step [17]. For that reason, here, we used a novel in-house bioinformatic pipeline that showed previously a remarkable performance in clustering otherwise un-clusterable complex disease to filter the publicly available data generated by transcriptomics approach. We aimed to identify novel pathways that can help in selecting potential biomarkers and explain the susceptibility to disease, predict prognosis, and response to treatment in a personalized medicine approach.

II. HYPOTHESIS

Proper image analysis of gene expression microarray can improve the identification of genuine differential expressed genes between severe asthma and healthy controls.

A. Aim

Use the publicly available gene expression data of patients with clinically distinct asthmatic phenotypes to identify a novel subset of genes correlated with disease severity (Healthy and Severe asthma) that can provide biological information on disease mechanisms and the pathways in which they are significantly enriched.

B. Objectives

1. Collect the raw image data from publicly available gene expression data of asthma patients.
2. Cluster patients with clinically distinct asthmatic phenotypes and identify confounding factors (age, sex, and sample type).
3. Use the method described by Hamoudi et al. [18] to identify a novel subset of genes correlated with disease severity.
4. Reveal how multiple mechanisms interact in asthma.
5. Construct a practical and efficient short-listed gene signature for different asthma phenotypes biomarkers.

III. METHODOLOGY AND RESULTS

110 records on Asthma with 7390 samples were searched in the publicly available transcriptomic dataset from Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). The datasets were filtered to search for a dataset that includes asthmatic patients with defined clinical classifications of participants, and bronchial epithelium gene expression using microarray

Affymetrix Human Genome U133 Plus 2.0 Array. Human Genome U133 Set plus have the advantages of complete coverage of over 47,000 transcripts plus the 6,500 additional genes for analysis. For initial identification of gene signature that differentiates healthy from severe asthma bronchial epithelium, the gene dataset GSE64913 [Healthy (n=37) and Severe Asthma (n=22)] was used to generate the signature. The general flowchart used in processing this dataset is detailed in Figure (1).

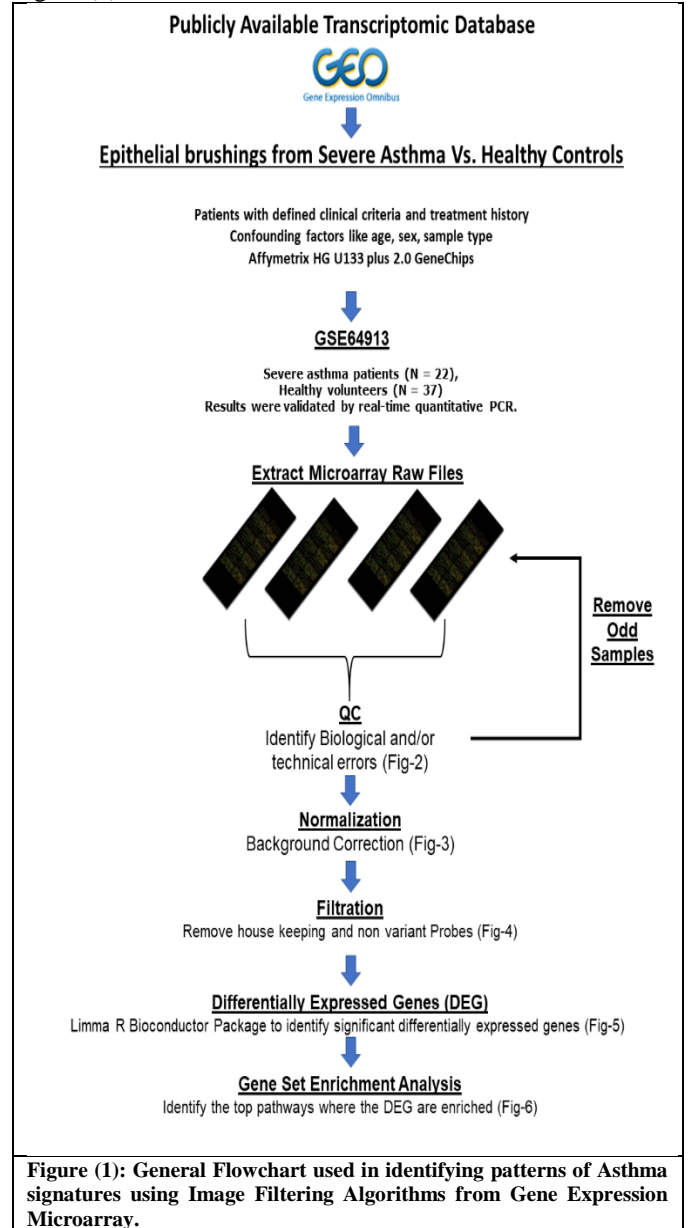


Figure (1): General Flowchart used in identifying patterns of Asthma signatures using Image Filtering Algorithms from Gene Expression Microarray.

A. Preprocessing Quality Control

Before submitting the datasets for differential expression analysis, the quality of the chips used for the study was assessed as per the manufacturer instructions, and only chips that passed the QC were used for further processing. Quality Control workflow is shown in Figure (2).

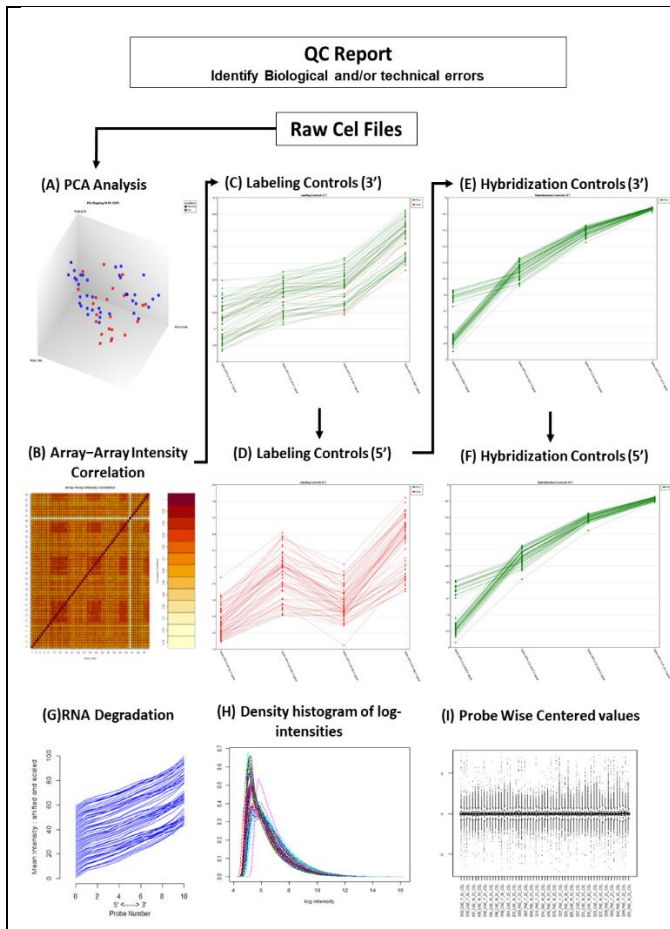


Figure (2): QC report to identify technical and biological errors that affect further processing generated by Transcriptome Analysis Console (TAC) 4.0 for microarray analysis and R Bioconductor package (affyQCReport). The report confirms that the 59 sample chips are appropriate for further analysis. (A) The PCA analysis showing the asthmatic samples in red and healthy samples in blue and there is no observed effect that can affect further interpretation (B) Array to array analysis for the 59 chips used in this study, showing dataset homogeneity (C-F) Set of probes spiked at the beginning of the chip processing to assess the overall success of the target preparation steps (G) RNA degradation plot to assess overall RNA quality control and whether RNA is too degraded or not (H) Density histogram of log-intensities for comparison between arrays and for an identification of arrays with weird distribution (I) Boxplot of the probe-wise centered values.

B. Pipeline for normalization

In total (n=59) with Affymetrix Human Genome U133 Plus 2.0 Array were extracted then underwent preprocessing and normalization separately using in-house pipeline using R Bioconductor to precise cluster otherwise unclusterable cases of one of the most biologically complex human disease using transcriptomic data as per the flowchart below. The pipeline uses affy package, Robust Multi-array Average (RMA), GCRMA, MAS5, and Li-Wong “dChip” of Bioconductor, R statistical software version 3.6.1. After testing the 4 methods, Hamoudi et al. [9] found that the best strategy for generating variant probes that are differentially expressed between the groups was found to be gcRMA and MAS5 and cross-reference the probes that passed through non-specific filtering based on

the coefficient of variation and absolute value thresholding. Flow chart of normalization and results are shown in figure (3).

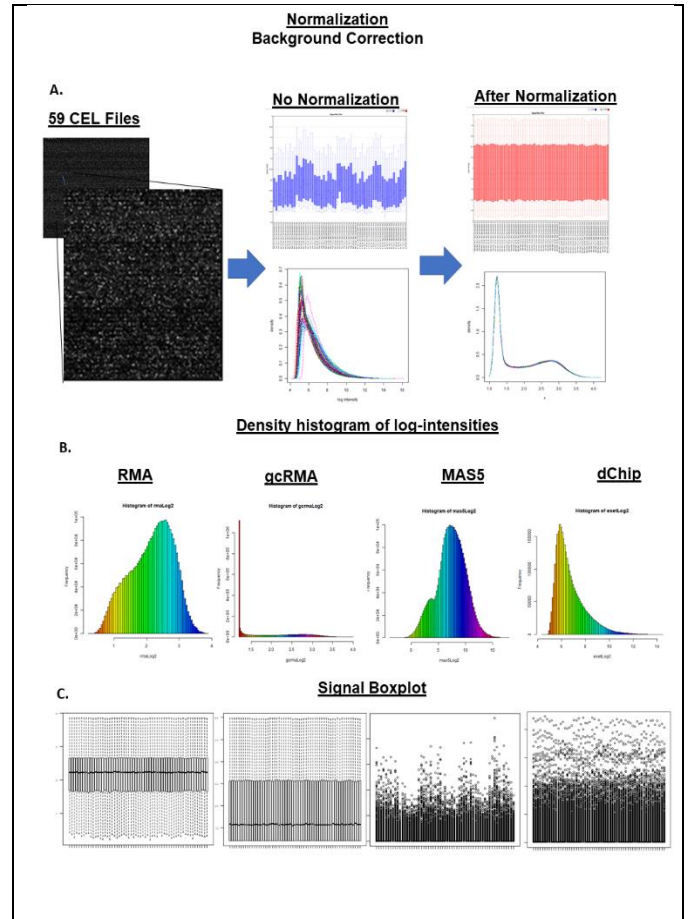
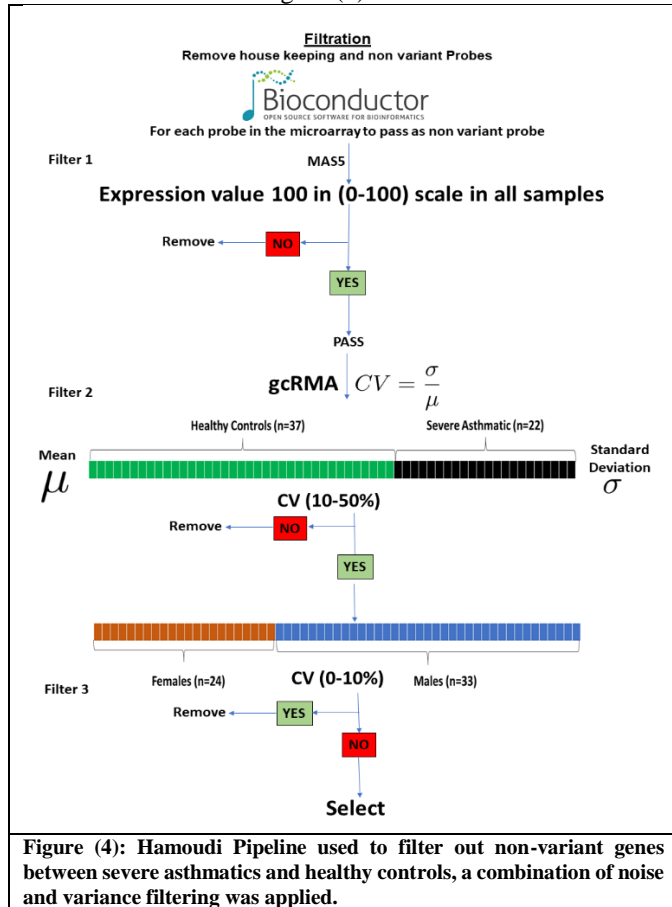


Figure (3): Different Normalization steps of removing sources of variation (backgrounds noise) which affect the measured gene expression levels using four commonly used normalization procedures RMA, GCRMA, MAS5, and Li-Wong “dChip”. Signal box plot for and density histogram of log intensities of each method are shown.

C. Pipeline for Filtration

To filter out non-variant genes between severe asthmatics and healthy controls, a combination of noise and variance filtering was applied. Only probes with a value of 100 or higher in the MAS5 dataset in all 59 samples were selected. The probes that passed the first filter than are subjected to the coefficient of variation (CV) filter using their gcRMA expression intensities. Probes with CV value of 10-50% across all samples were considered to be variant and thus selected. CV was calculated as the mean / standard deviation of each gene across all samples. Since many genes are different between males and females and should be identified, only genes that do not show significant variance between males and females samples were passed. Finally, the genes that passed the above three filtering methods were intersected to obtain a common set of variant genes. Such an approach will give us a clear list of genes that participate in the pathogenesis of severe asthma. Out of the 54675 probes present in the chip, only 3098 probes passed the filtration. These filtered probes were annotated to its

corresponding genes using (biomaRt) R package and then collapsed using the GSEA tool by choosing probes with the maximum expression for each gene. Initial 2474 protein-coding, 290 long RNA coding, and 9 microRNA coding genes were collapsed into 2116 corresponding genes. The flow chart of filtration is shown in figure (4).



D. Limma Package to identify DEG

To identify differentially expressed genes between severe asthma and healthy controls, the R Bioconductor Limma package was used. Out of the 2116 filtered genes, 169 genes with an adjusted p-value of less than 0.05 were identified to be differentially expressed between severe asthma and healthy controls. The flow chart of filtration is shown in figure (5).

Out of these, 16 genes (FKBP5, KRT23, HEG1, S100A10, SLAMF7, TFCP2L1, KLRC4-KLRK1, TRIM7, SLC13A2, SCGB3A1, C3, TM4SF1, PIP, NELL2, HLA-DPA1, and HLA-DQA2) showed >1 or <-1-fold change between the two groups as shown in table (1). These genes are important in the innate and adaptive immune response against microbial infections and regulating the cytokine production in health and autoimmune diseases.

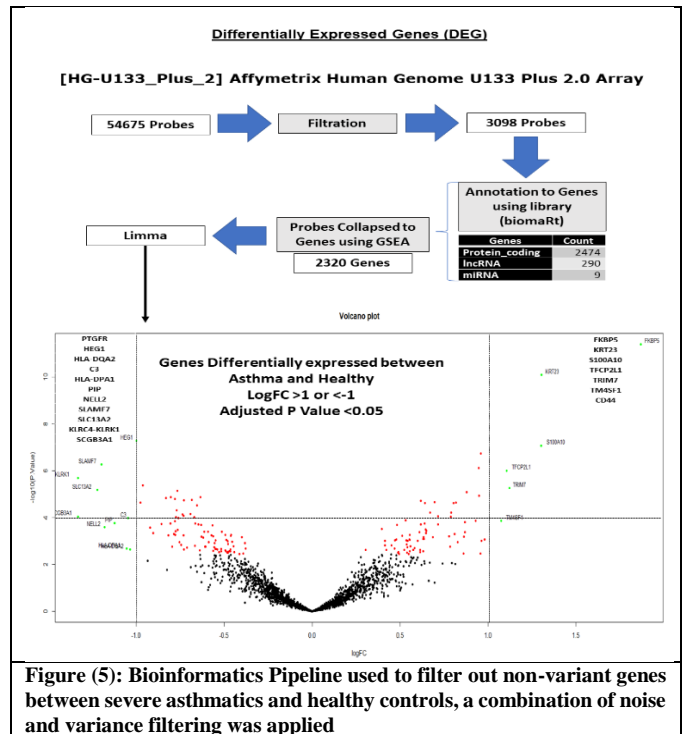


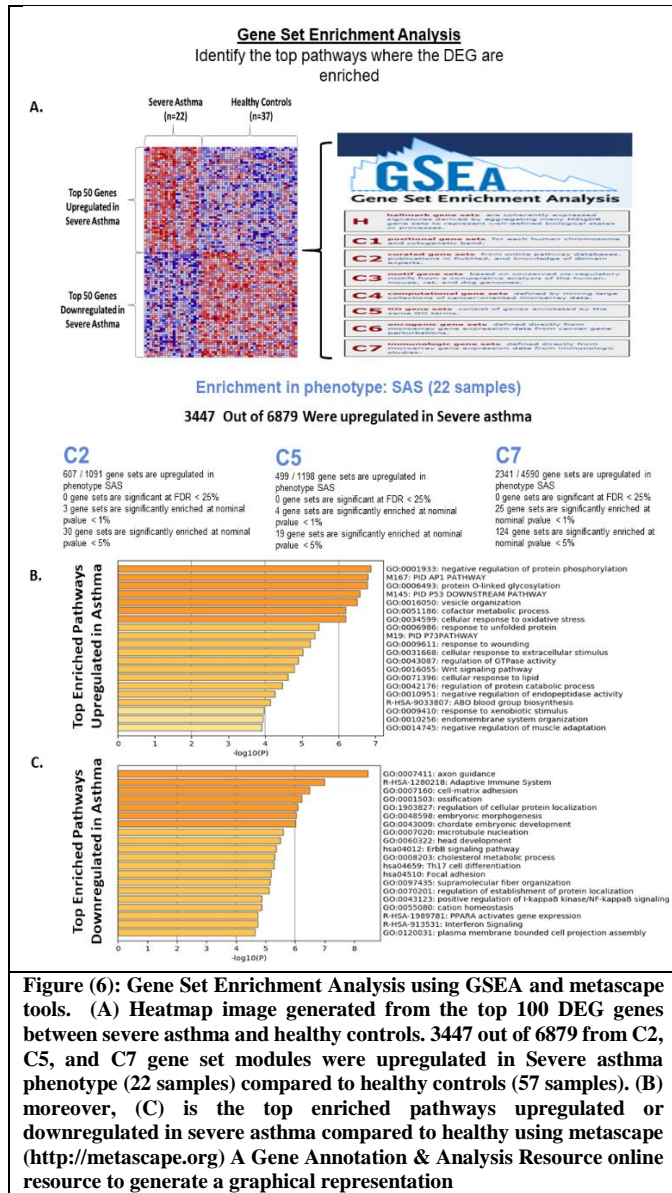
Table (1): Top differentially expressed Genes between Severe asthmatic and healthy controls bronchial epithelium.

Gene	Description	logFC	P.Value	adj.P.Val
FKBP5	FKBP prolyl isomerase 5	1.868512	3.94E-12	1.22E-08
KRT23	keratin 23	1.304043	8.06E-11	1.25E-07
S100A10	S100 calcium binding protein A10	1.301535	8.39E-08	4.33E-05
TRIM7	tripartite motif containing 7	1.121105	5.52E-06	0.001069
TFCP2L1	transcription factor CP2 like 1	1.105197	9.77E-07	0.000275
TM4SF1	transmembrane 4 L six family member 1	1.074957	0.000135	0.007411
HEG1	heart development protein with EGF like domains 1	-1.00061	5.22E-08	3.24E-05
HLA-DQA2	major histocompatibility complex, class II, DQ alpha 2	-1.03517	0.002303	0.038986
C3	complement C3	-1.04853	0.000104	0.006081
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	-1.05467	0.002118	0.037974
PIP	prolactin induced protein	-1.12388	0.00017	0.008474
NELL2	neural EGFL like 2	-1.18135	0.000255	0.010809
SLAMF7	SLAM family member 7	-1.19804	5.36E-07	0.000198
SLC13A2	solute carrier family 13 member 2	-1.22238	6.56E-06	0.001195
KLRC4-KLRK1	KLRC4-KLRK1 readthrough	-1.33116	2.02E-06	0.000523
SCGB3A1	secretoglobin family 3A member 1	-1.33309	8.96E-05	0.005854

E. Gene Set Enrichment Analysis

Gene set enrichment analysis was done as previously described [18] to identify top pathways in which the identified DEG are highly enriched to shed light on the molecular pathogenesis of asthma by deconvoluting the expression profiling signals generated from the same gene (datapoint) having different levels of expression based on the degree of Asthma severity. The flow chart of GSEA and details of enriched pathways are shown in figure (6). 3447 out of 6879 from C2, C5, and C7 gene set modules were upregulated in Severe asthma phenotype (22 samples) compared to healthy controls (57 samples). The Genes

that were upregulated or downregulated in severe asthma were uploaded separately to metascape (<http://metascape.org>) A Gene Annotation & Analysis Resource online resource to generate a graphical representation of top enriched pathways.

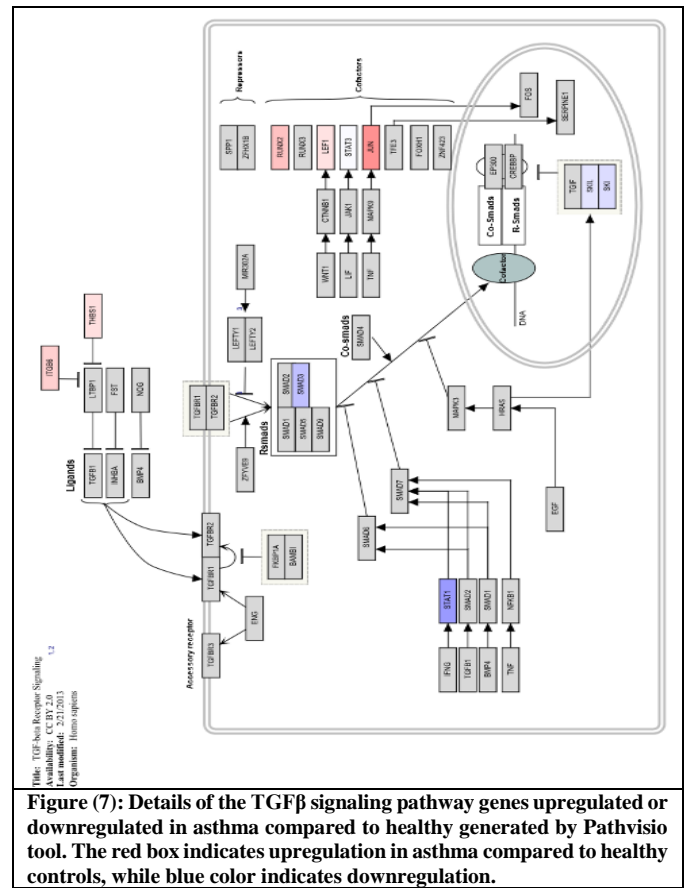


F. Members of Interesting Pathways (TGFβ and NFκB) were identified to be preferentially expressed in the asthmatic epithelium

Gene members of two interesting pathways were identified by our pipeline, namely Transforming growth factor β (TGFβ) and nuclear factor kappa-light-chain-enhancer of activated B cells (NFκB) to be preferentially expressed in the asthmatic epithelium. We used PathVisio open-source biological pathway analysis software to draw and analyze biological pathways by uploading the DEG with their log fold change between asthma and healthy controls.

TGFβ

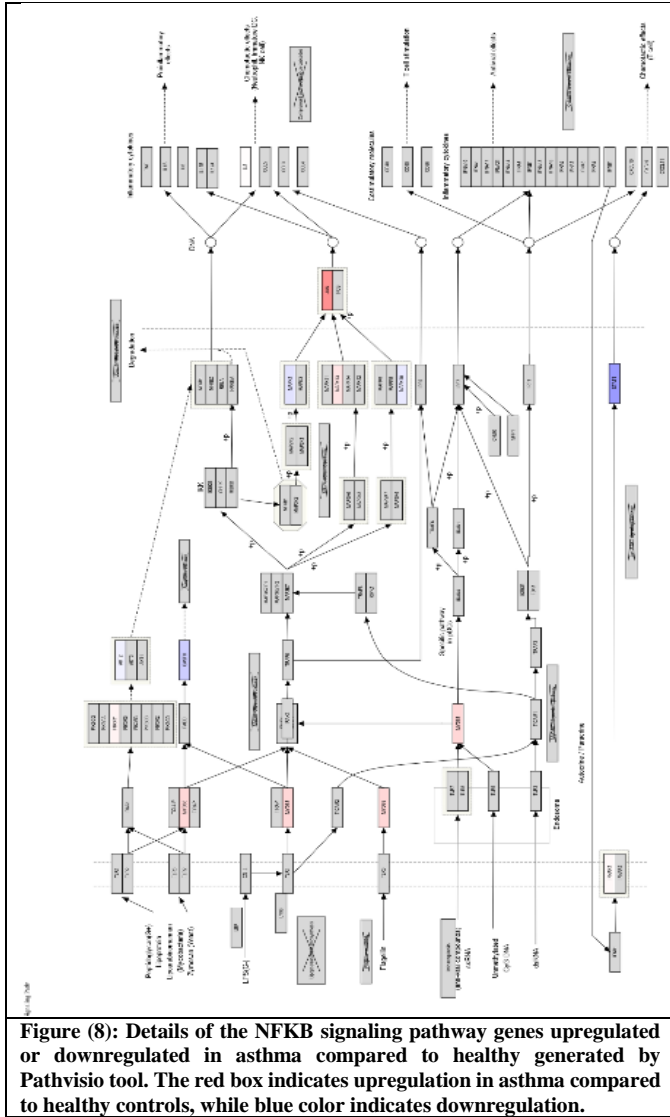
TGFβ is a critical regulator of pro-inflammatory responses and fibrotic tissue remodeling within the asthmatic lung [19]. Our analysis identified that among the upregulated TGFβ pathway members are ITGB6 and THBS1 genes which regulate TGFβ release from regulatory Latency-associated peptide (LAP), thereby playing a pivotal role in TGFβ activation. Furthermore, two important cofactors for SMADs DNA binding to exert the TGFβ1 effect (JUN and RUNX2) were upregulated. On the other hand, TGFβ signaling repressors (SKIL and Ski) were downregulated in severe asthma compared to the healthy epithelium, confirming an activated TGFβ signaling. Figure (7) shows the details of the TGFβ signaling pathway genes upregulated or downregulated in asthma compared to healthy generated by Pathvisio tool.



NFκB

NFκB is a critical transcription factor activated in the airway epithelium of human asthmatics and mice after allergic stimulation and control production of many inflammatory cytokines [20]. Our results showed that MYD88 (Myeloid differentiation primary response 88 genes) was upregulated in severe asthma compared to healthy controls. MYD88 is a universal adapter protein used by Toll-like receptors (TLRs) which are a class of proteins that play a vital role in the innate immune system by activating NF-κB. Downstream of MYD88, MAPK13, and JUN are upregulated in asthma to stimulate IL12B and CCL5. IL12 is a cytokine that acts on T and natural

killer cells and was linked to asthma in children. On the other hand, CCL5 is chemotactic for T cells, eosinophils, and basophils into inflammatory sites. Figure (8) shows the details of the NFKB signaling pathway genes upregulated or downregulated in asthma compared to healthy generated by Pathvisio tool.



IV. CONCLUSIONS

Our analysis identified important signaling pathways members to be differentially expressed between severe asthma and healthy controls, namely NFKB and TGF β pathways. Those vital members represent potential targets for future asthma treatment and can serve as a reliable biomarker for asthma severity. Also, this study showed that using image analysis methodology outlined in the study on publicly available microarray transcriptomics data can be used to decipher the molecular mechanism of complex diseases such as asthma by deconvoluting the similar signals generated from the gene expression profiles.

V. ACKNOWLEDGMENT

We thank the Sharjah Research Academy (Grant code: MED001) and Al-Jalila Foundation (Grant code: AJF201741) for funding this work.

REFERENCES

- [1]. Braman, S.S., The global burden of asthma. *Chest*, 2006. 130(1 Suppl): p. 4S-12S.
- [2]. Field, J.J., et al., Airway Hyperresponsiveness in Children With Sickle Cell Anemia. *Chest*, 2011. 139(3): p. 563-568.
- [3]. Elina, T. and K.D. W., Asthma risk factors. *International Forum of Allergy & Rhinology*, 2015. 5(S1): p. S11-S16.
- [4]. Vianello, A., et al., Fatal asthma; is it still an epidemic? *The World Allergy Organization Journal*, 2016. 9(1): p. 42.
- [5]. P, H.R. and K.L. C, 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 2011. 13(3): p. 189-195.
- [6]. Galeone, C., et al., Precision Medicine in Targeted Therapies for Severe Asthma: Is There Any Place for "Omics" Technology? *BioMed research international*, 2018. 2018: p. 4617565-4617565.
- [7]. Barnes, K.C., Genetic studies of the etiology of asthma. *Proceedings of the American Thoracic Society*, 2011. 8(2): p. 143-148.
- [8]. Dahlin, A. and K.G. Tantisira, Integrative systems biology approaches in asthma pharmacogenomics. *Pharmacogenomics*, 2012. 13(12): p. 1387-1404.
- [9]. Persson, H., et al., Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles. *Journal of Allergy and Clinical Immunology*, 2015. 136(3): p. 638-648.
- [10]. Stewart, C.E., et al., Evaluation of Differentiated Human Bronchial Epithelial Cell Culture Systems for Asthma Research. *Journal of Allergy*, 2012. 2012: p. 11.
- [11]. Kuo, C.S., et al., A Transcriptome-driven Analysis of Epithelial Brushings and Bronchial Biopsies to Define Asthma Phenotypes in U-BIOPRED. *Am J Respir Crit Care Med*, 2017. 195(4): p. 443-455.
- [12]. Modena, B.D., et al., Gene expression in relation to exhaled nitric oxide identifies novel asthma phenotypes with unique biomolecular pathways. *Am J Respir Crit Care Med*, 2014. 190(12): p. 1363-72.
- [13]. Kim, R.D. and P.J. Park, Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome biology*, 2004. 5(9): p. R70-R70.
- [14]. Bryant PA, Smyth GK, Robins-Browne R, Curtis N. Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation. *PLoS One*. ;6(5):e19556. doi:10.1371/journal.pone.0019556
- [15]. Calza, S. and Y. Pawitan, Normalization of gene-expression microarray data. *Methods Mol Biol*, 2010. 673: p. 37-52.
- [16]. Bengtsson, A. and H. Bengtsson, Microarray image analysis: background estimation using quantile and morphological filters. *BMC Bioinformatics*, 2006. 7(1): p. 96.
- [17]. Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*. 2009;10:11. Published 2009 Jan 8. doi:10.1186/1471-2105-10-11
- [18]. Hamoudi, R.A., et al., *Differential expression of NF-kappaB target genes in MALT lymphoma with and without chromosome translocation: insights into molecular mechanism*. *Leukemia*, 2010. 24(8): p. 1487-97.
- [19]. Al-Alawi, M., T. Hassan, and S.H. Chotirmall, Transforming growth factor β and severe asthma: A perfect storm. *Respiratory Medicine*, 2014. 108(10): p. 1409-1423.

Sheller, J.R., et al., Nuclear factor kappa B induction in airway epithelium increases lung inflammation in allergen-challenged mice. *Experimental lung research*, 2009. 35(10): p. 883-895.

