Identifying Common Molecular Signatures between Severe Asthma and Lung Cancer

Aditi Sarathy
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
aditisarathy@cmu.edu

Mirudhula Mukundan
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
mirudhum@andrew.cmu.edu

1 Introduction

Lung cancer continues to be the most severe form of cancer being the leading cause of all cancer-related death worldwide. More than half the number of patients are known to have advanced stages of lung cancer by the time they are diagnosed. Incidence of lung cancer is higher in people over 60 years and above, it can occur due to environmental factors, hereditary or smoking(13). Common perception indicates smoking to be the most likely reason but that does not explain for the 25% of the cases attributed to non-smoking related lung cancer(23).

Asthma, Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis and so on are known to be pulmonary co-morbidities related to lung cancer and the presence of these co-morbidities has shown to result in early diagnosis of cancers (6). There are many factors that can cause chronic inflammation in the bronchial epithelium which can result in lung cancer. There are also studies which have shown that inflammatory state in severe asthma can make patients susceptible to cancer of lung and other organs (20). Some meta-studies have linked severe cases of asthma to lung cancer (21; 17; 19), but these studies are mostly associated with conducting risk analysis on asthmatic and cancer patients, being observed over the course of several years, indicating a dearth in research based on gene expression data.

Although there are studies which have identified the molecular signatures and associated pathways of asthma (2) and lung cancer (25) independently, very few (20) have reported key molecular signatures associated with both lung cancer and severe case of asthma. Hence, in this project, we identified a panel of gene signatures that are associated in both Non Small-cell Lung Cancer (NSCLC) patients and Severe Asthmatic (SA) patients, through insilico methods using the gene expression data of both the diseases in epithelial cells of the bronchial tract. We validated these genes and arrived at 8 key genes that seemed to be differentially expressed in Lung Cancer (overexpressed) and Severe Asthma (underexpressed). Out of them, we identitfied PPARD to be expressed in higher levels in mixed cases of the diseases. This reveals that these genes could possibly be identified as biomarkers of NSCLC in patients with severe cases of asthma.

2 Data

2.1 Dataset

For analysis, gene expression data for Severe Asthma (GSE64913) and NSCLC (GSE29013) was downloaded from Gene Expression Omnibus. The asthma dataset used was obtained from epithelial brushings of peripheral airways of the patients. This dataset consists of 28 severe asthmatic patients and 42 healthy volunteers and a total expression data from 54675 genes. Expression data for NSCLC was obtained from Formalin-Fixed Paraffin-Embedded Samples (FFPE), which is known to be a good

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

source to study the molecular changes in cancer and the associated clinical outcome. This dataset contained 55 samples of patients with lung cancer and expression data from 54675 genes. Out of the 55 NSCLC patients, only the 52 non-smoking patients were considered. The sequences were analyzed on Affymetrix microarrays to obtain the expression data.

The results obtained were validated on another pair of gene sets for Asthma (GSE63142) and NSCLC (GSE68793). The gene expression data for asthma was collected from bronchial epithelial cells of asthma, of 155 samples in total out of which only 36 samples labeled as "Severe Asthmatic", was used for this study. Similarly, out of the 135 NSCLC patients, only 39 patients who were mentioned to be non-smokers or reformed smokers for more than 15 years was chosen as the final cohort.

2.2 Data Preprocessing

Normalized Gene expression data for Severe Asthma and Non Small Cell Lung Cancer (NSCLC) was available from Gene Expression Omnibus. The Asthma gene expression data had labels for both healthy and severe asthmatic samples, and no features with zero/NaN values were found. The Lung cancer gene expression data which only had samples with lung cancer patients was combined with the normal patient samples from the asthma dataset, thereby using the same control samples across both diseases. Additonally, those who had zero counts in a particular feature was removed.

3 Methods

The workflow is depicted in the flowchart in Fig.1.

3.1 Differential Gene Expression

After collecting the gene expression datasets for Severe Asthma and Lung cancer, Differential Gene Expression Analysis (DGE) was performed to determine which genes are expressed at different levels between disease and healthy samples. This was done using R programming language, utilizing the Limma package in Bioconductor (7). The up-regulated and down-regulated genes for both these conditions were identified.

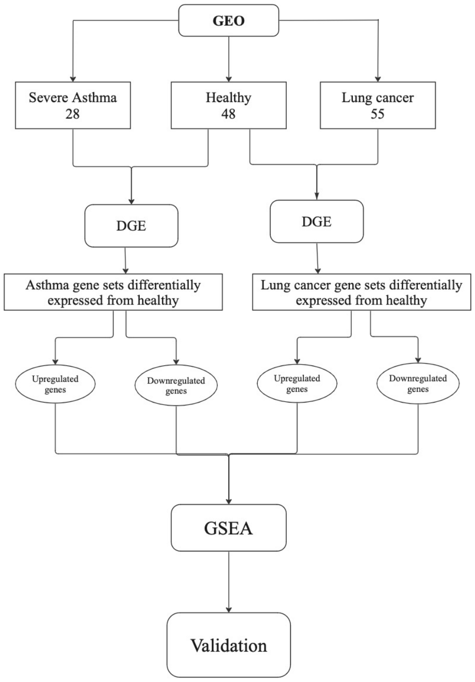


Figure 1: Overview of Methodology

3.2 Gene Set Enrichment Analysis

Following DGE, we were able to form four different gene sets, that could be used for performing Gene Set Enrichment Analysis (GSEA) (22). This was done to analyze two aspects: (a) GSEA of upregulated/downregulated lung cancer gene sets in asthma dataset, (b) GSEA of upregulated/downregulated asthma gene sets in lung cancer dataset. Genes were ranked in each case by computing their correlation to the disease and control samples, and ordered accordingly. Gene shuffling was conducted with 1000 permutations, which allowed estimation of p-values and false discovery rate with a precision of upto 10^{-3}. The leading edge set representing the top genes that are enriched in each experiment of GSEA was identified. This helped us identifying the expression pattern of, for example, a gene set that is upregulated in lung cancer, in asthma dataset. This was done in order to select only those genes that seem to be overrepresented in the dataset the gene set is matched against.

3.3 Validation

To validate the set of genes that we obtained from GSEA, we intersected these with another pair of asthma-NSCLC datasets to identify common features which can be validated. Then, we implemented a decision tree learning model on MATLAB to see if the genes were expressed similar to our analysis datasets. This methodology followed is quite similar to the one used by Irshad *et al.* (11). The gene expressions were z-score normalized across all samples, and the values were further discretized as: (a) 1, if e_i >= 0.5, (b) -1, if e_i <= -0.5, and (c) 0, otherwise. These represent the expression states of each gene with respect to each patient, and a series of decision trees were implemented using this data, for an increasing number of features for all combinations of genes, i.e. d-trees construction was iteratively done by adding more genes to the predictive combinations. Finally, tree pruning was also conducted to avoid overfitting. The losses for each tree was calculated and the best combination of gene was selected based on minimal loss in the test set, after conducting a 5-fold cross validation.

4 Results

4.1 Differential Gene Expression Analysis

DGE was performed on R studio by using *limma* package. To improve the accuracy of the prediction we filtered the lowly expressed genes and selected only those genes which should a high expression using an absolute log fold change cut off. Further only those genes were chosen that had a high expression for at least 2 samples. The top genes were ranked having an absolute log fold change (FC) and p-value cut off chosen for the two datasets shown in Table 1.

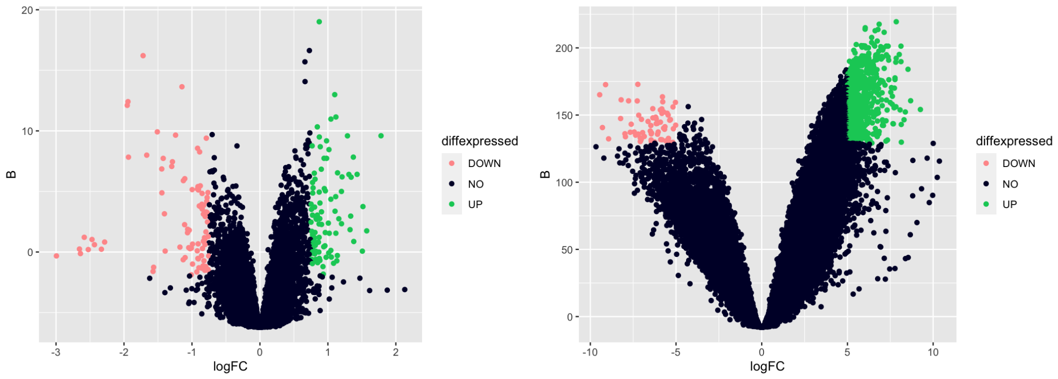
Table 1: Abs log fold change and p-value

Gene Set	Pvalue	$abs(\log_2(FC))$
Asthma	$10e^{-60}$	5
Lung Cancer	$5e^{-2}$	0.75

Genes with a negative $abs(\log_2(FC))$ were considered to be the down-regulated genes and genes with a positive $abs(\log_2(FC))$ were considered to be up-regulated (See Table 2, for reference). A volcano plot indicating the up-regulated and down-regulated genes for both severe asthma and lung cancer was obtained as shown in Fig.2. These gene categories form the four different gene sets that were used in GSEA.

Table 2: Up-regulated and Down-regulated genes for Severe Asthma and Lung Cancer

Condition	Up-regulated	Down-regulated
Asthma	98	87
Lung Cancer	950	190



(a) DGE for Asthma vs Healthy

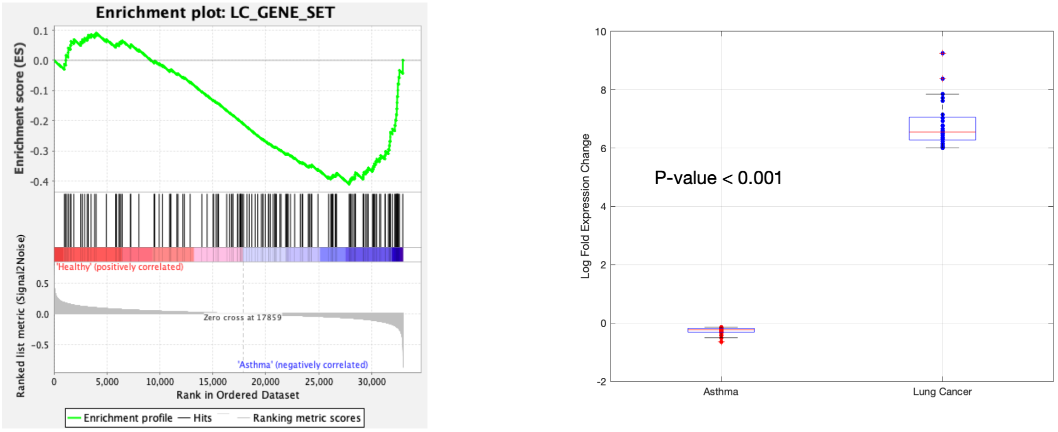
(b) DGE for Lung Cancer vs Healthy

Figure 2: Identifying the up-regulated and down-regulated genes for severe asthma and lung cancer: (a) Severe Asthma, (b) Lung Cancer

4.2 Gene Set Enrichment Analysis

Our goal was to observe whether the filtered differentially expressed lung cancer gene sets were enriched in the Asthma dataset and vice versa. As shown in Fig.3a, enrichment of upregulated lung cancer gene set in the asthma dataset was observed with a p-value less than 0.05 and having a False Discovery Rate (FDR) equal to 5%. Leading edge set of 40 genes were identified that were indicated to be overrepresented in the asthma dataset. In the other three GSEA analysis, no significant results were obtained. The downregulated genes of lung cancer seemed to have a uniform representation in the asthma gene set, and produced a p-value > 0.05. There appeared to be some enrichment of upregulated asthma genes in the NSCLC dataset, but almost all genes were highly correlated with control patients and not asthma patients. Likewise, downregulated genes of asthma had uniform representation with respect to both NSCLC and control patients, so these results had to eliminated as well, for further analysis (See Fig. S1a, S1b, S1c).

For the 40 top enriched genes selected from GSEA against asthma dataset, we identifies their p-values in the same dataset, to see if the expression was significant. We filtered it down to 25 genes that had significance less than 0.05, and removed genes from this leading edge set for which we could not infer gene symbols, leaving us with 22 significant genes. Furthermore, we found that these 22 genes were, in fact, being differentially expressed in both NSCLC and Severe Asthma patients, wherein the genes were under-expressed in asthma but overexpressed in NSCLC (Refer Fig.3b).



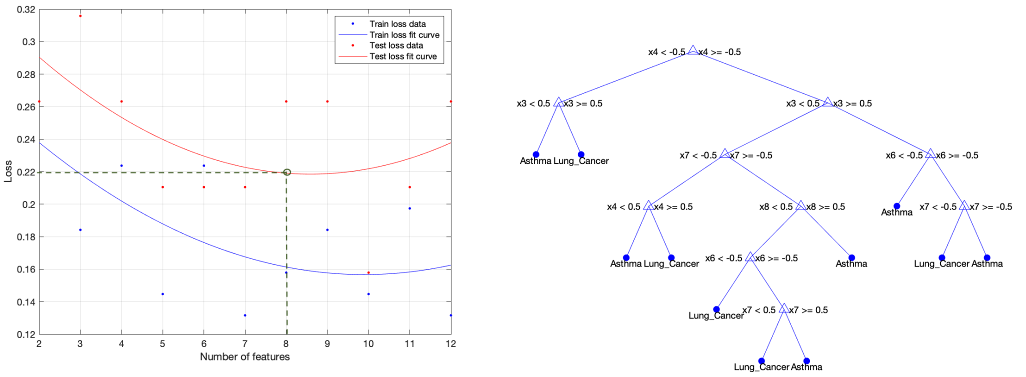
(a)

(b)

Figure 3: (a) GSEA for up-regulated gene set of lung cancer against severe asthma dataset, (b) Log fold expression change for Asthma and Lung cancer. Boxplot shows differential expression in both diseases

4.3 Validation

The top 22 filtered genes were intersected with another pair of asthma-lung cancer datasets, to get 15 common genes to be validated upon. Decision tree was implemented to classify NSCLC and Severe Asthma, iteratively for an increasing number of genes or features, and the loss for the best permutation of genes was plotted for each iteration (Fig.4a). From this result, a list of final 8 genes (AAK1, CALD1, HIF1A, KIAA0101, PPARD, PPP1R13L, SCRIB, SIN3B) were identified to have the best classification accuracy, without overfitting the training data.



(a)

(b)

Figure 4: (a) Train-Test loss plot from implementing decision classification tree, (b) Final Decision Tree with 8 features

5 Discussion & Conclusion

This study helped us identify 8 key molecular signatures in differentially expressed genes of Asthma and Lung cancer. All 8 genes were somehow related to lung cancer or asthma. The complete details about the 8 genes are listed in Supplementary Table S1. PCLAF was found to be involved in signaling and DNA repair mechanism pathways, which is in fact related to 2.5% of adenocarcinoma cases are due to mutations in DNA repair genes (3). This gene was also found to be responsible for orchestrating major inflammatory problems in airspaces (16). It was surprising to find this genebeing differentially expresses in two diseases that are both related to respiratory tract. Furthermore, PPARD and CALD1 (Caldesmon) are known to have functional roles in adhesion, inflammation, proliferation and regulating interactions in smooth muscles. Additionally, both are known to be markers for both asthma and lung cancer (1; 30). Another interesting aspect of PPARD is that it was observed to have an elevated gene expression, compared to plain asthma cases, in patients having mixed cases of severe asthma and NSCLC. We hypothesize that this gene can be a key factor enabling us to use it as a potential marker for diagnosing early lung cancer in severe asthma patients. As next steps, conducting *in vivo* analysis of this 8 gene panel would possibly help uncover more relations with respect to severe asthma, NSCLC and the mixed cases of these diseases.

References

[1] ALNUAIMI, A. R., NAIR, V. A., MALHAB, L. J. B., ABU-GHARBIEH, E., RANADE, A. V., PINTUS, G., HAMAD, M., BUSCH, H., KIRFEL, J., HAMOUDI, R., ET AL. Emerging role of caldesmon in cancer: A potential biomarker for colorectal cancer and other cancers. *World Journal of Gastrointestinal Oncology 14*, 9 (2022), 1637.

[2] BIGLER, J., BOEDIGHEIMER, M., SCHOFIELD, J. P., SKIPP, P. J., CORFIELD, J., ROWE, A., SOUSA, A. R., TIMOUR, M., TWEHUES, L., HU, X., ET AL. A severe asthma disease signature from gene expression profiling of peripheral blood from u-biopred cohorts. *American journal of respiratory and critical care medicine 195*, 10 (2017), 1311–1320.

[3] Burgess, J. T., Rose, M., Boucher, D., Plowman, J., Molloy, C., Fisher, M., O'Leary, C., Richard, D. J., O'Byrne, K. J., and Bolderson, E. The therapeutic potential of dna damage repair pathways and genomic stability in lung cancer. *Frontiers in Oncology 10* (2020), 1256.

[4] Chen, Y., Pan, Y., Ji, Y., Sheng, L., and Du, X. Network analysis of differentially expressed smoking-associated mrnas, lncrnas and mirnas reveals key regulators in smoking-associated lung cancer. *Experimental and therapeutic medicine 16*, 6 (2018), 4991–5002.

[5] Conner, S. D., and Schmid, S. L. Identification of an adaptor-associated kinase, aak1, as a regulator of clathrin-mediated endocytosis. *The Journal of cell biology 156*, 5 (2002), 921–929.

[6] Dima, S., Chen, K.-H., Wang, K.-J., Wang, K.-M., and Teng, N.-C. Effect of comorbidity on lung cancer diagnosis timing and mortality: A nationwide population-based cohort study in taiwan. *BioMed research international 2018* (2018).

[7] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology 5*, 10 (2004), 1–16.

[8] Han, F., Chen, G., Guo, Y., Li, B., Sun, Y., Qi, X., Tian, H., Zhao, X., and Zhang, H. Microrna-4491 enhances cell proliferation and inhibits cell apoptosis in non-small cell lung cancer via targeting trim7. *Oncology Letters 22*, 2 (2021), 1–9.

[9] Hu, S., Zeng, W., Zhang, W., Xu, J., Yu, D., Peng, J., and Wei, Y. Kiaa0101 as a new diagnostic and prognostic marker, and its correlation with gene regulatory networks and immune infiltrates in lung adenocarcinoma. *Aging (Albany NY) 13*, 1 (2021), 301.

[10] Huang, G.-H., Zhang, Y.-H., Chen, L., Li, Y., Huang, T., and Cai, Y.-D. Identifying lung cancer cell markers with machine learning methods and single-cell rna-seq data. *Life 11*, 9 (2021), 940.

[11] Irshad, S., Bansal, M., Castillo-Martin, M., Zheng, T., Aytes, A., Wenske, S., Le Magnen, C., Guarnieri, P., Sumazin, P., Benson, M. C., et al. A molecular signature predictive of indolent prostate cancer. *Science translational medicine 5*, 202 (2013), 202ra122–202ra122.

[12] Kunneman, M., Marijnen, C. A., Rozema, T., Ceha, H. M., Grootenboers, D. A., Neelis, K. J., Stiggelbout, A. M., and Pieterse, A. H. Decision consultations on preoperative radiotherapy for rectal cancer: large variation in benefits and harms that are addressed. *British journal of cancer 112*, 1 (2015), 39–43.

[13] Lee, K., Kang, S., and Hwang, J. Lung cancer patients' characteristics and comorbidities using the korean national hospital discharge in-depth injury survey data. *Journal of Epidemiology and Global Health* (2022), 1–9.

[14] Lewis, M. J., Liu, J., Libby, E. F., Lee, M., Crawford, N. P., and Hurst, D. R. Sin3a and sin3b differentially regulate breast cancer metastasis. *Oncotarget 7*, 48 (2016), 78713.

[15] Liu, J., and Liu, X. Ube2t silencing inhibited non-small cell lung cancer cell proliferation and invasion by suppressing the wnt/β-catenin signaling pathway. *International Journal of Clinical and Experimental Pathology 10*, 9 (2017), 9482.

[16] Mould, K. J., Jackson, N. D., Henson, P. M., Seibold, M., and Janssen, W. J. Single cell rna sequencing identifies unique inflammatory airspace macrophage subsets. *JCI insight 4*, 5 (2019).

[17] Qu, Y.-L., Liu, J., Zhang, L.-X., Wu, C.-M., Chu, A.-J., Wen, B.-L., Ma, C., Yan, X.-y., Zhang, X., Wang, D.-M., et al. Asthma and the risk of lung cancer: a meta-analysis. *Oncotarget 8*, 7 (2017), 11614.

[18] Reddy, A. T., Lakshmi, S. P., and Reddy, R. C. Pparγ as a novel therapeutic target in lung cancer. *PPAR research 2016* (2016).

[19] Rosenberger, A., Bickeboeller, H., McCormack, V., Brenner, D. R., Duell, E. J., Tjønneland, A., Friis, S., Muscat, J. E., Yang, P., Wichmann, H.-E., et al. Asthma and lung cancer risk: a systematic investigation by the international lung cancer consortium. *Carcinogenesis 33*, 3 (2012), 587–597.

[20] Salameh, L., Bhamidimarri, P. M., Saheb Sharif-Askari, N., Dairi, Y., Hammoudeh, S. M., Mahdami, A., Alsharhan, M., Tirmazy, S. H., Rawat, S. S., Busch, H., et al. In silico bioinformatics followed by molecular validation using archival ffpe tissue biopsies identifies a panel of transcripts associated with severe asthma and lung cancer. *Cancers 14*, 7 (2022), 1663.

[21] Salameh, L., Mahboub, B., Khamis, A., Alsharhan, M., Tirmazy, S. H., Dairi, Y., Hamid, Q., Hamoudi, R., and Al Heialy, S. Asthma severity as a contributing factor to cancer incidence: A cohort study. *PloS one 16*, 5 (2021), e0250430.

[22] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences 102*, 43 (2005), 15545–15550.

[23] Sun, S., Schiller, J. H., and Gazdar, A. F. Lung cancer in never smokers—a different disease. *Nature Reviews Cancer 7*, 10 (2007), 778–790.

[24] Swinson, D. E., Jones, J. L., Cox, G., Richardson, D., Harris, A. L., and O'Byrne, K. J. Hypoxia-inducible factor-1α in non small cell lung cancer: Relation to growth factor, protease and apoptosis pathways. *International journal of cancer 111*, 1 (2004), 43–50.

[25] Taniwaki, M., Daigo, Y., Ishikawa, N., Takano, A., Tsunoda, T., Yasui, W., Inai, K., Kohno, N., and Nakamura, Y. Gene expression profiles of small-cell lung cancers: molecular signatures of lung cancer. *International journal of oncology 29*, 3 (2006), 567–575.

[26] Wang, X., Li, G., Koul, S., Ohki, R., Maurer, M., Borczuk, A., and Halmos, B. Phlda2 is a key oncogene-induced negative feedback inhibitor of egfr/erbb2 signaling via interference with akt signaling. *Oncotarget 9*, 38 (2018), 24914.

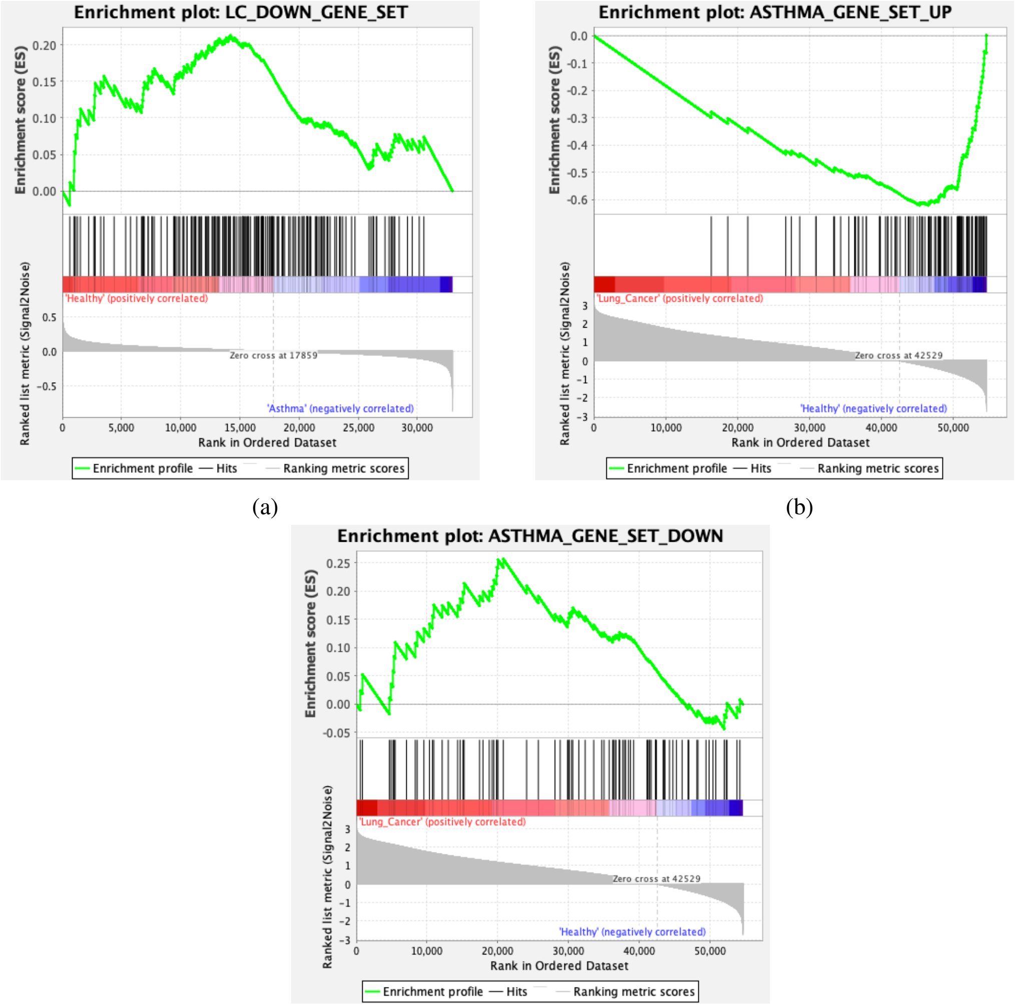
[27] Wang, Y., Liu, H., Bian, Y., An, J., Duan, X., Wan, J., Yao, X., Du, C., Ni, C., Zhu, L., et al. Low scrib expression in fibroblasts promotes invasion of lung cancer cells. *Life Sciences 256* (2020), 117955.

[28] Yang, J.-P., Hori, M., Sanda, T., and Okamoto, T. Identification of a novel inhibitor of nuclear factor-κb, rela-associated inhibitor. *Journal of Biological Chemistry 274*, 22 (1999), 15662–15670.

[29] Zhou, X., Sun, Q., Xu, C., Zhou, Z., Chen, X., Zhu, X., Huang, Z., Wang, W., and Shi, Y. A systematic pan-cancer analysis of pxdn as a potential target for clinical diagnosis and treatment. *Frontiers in Oncology 12* (2022).

[30] Zingarelli, B., Piraino, G., Hake, P. W., O'Connor, M., Denenberg, A., Fan, H., and Cook, J. A. Peroxisome proliferator-activated receptor δ regulates inflammation via nf-κb signaling in polymicrobial sepsis. *The American journal of pathology 177*, 4 (2010), 1834–1847.

6 Appendix



(a)

(b)

(c)

Figure S1: GSEA for up-regulated and down-regulated genes for severe asthma and lung cancer: (a) Down-regulated gene set of lung cancer against severe asthma dataset, (b) Up-regulated gene set of asthma against severe lung cancer dataset, (c) Down-regulated gene set of asthma against severe lung cancer dataset

Table S1: Details about the 8 key genes correlated with both NSCLC and Severe Asthma

Gene	Protein name	Pathway/Description	Reference
AAK1	AP2-associated protein kinase 1	Clathrin mediated endocytosis; Prognostic marker for ovarian cancer.	(5)
CALD1	Caldesmon	It is an actin and myosin-binding protein implicated in the regulation of actomyosin interactions in smooth muscle and nonmuscle cells. CaLD is known to be overexpressed in brain metastases of lung Cancer.	(1)
DNALI1	Axonemal dynein light intermediate polypeptide 1	May play a dynamic role in flagellar motility. Found to be downregulated in lung cancer in patients with a smoking history.	(4)
HIF1A	Hypoxia-inducible factor 1-alpha	Functions as a master transcriptional regulator of the adaptive response to hypoxia. HIF1A is commonly expressed in NSCLC and is associated with a number of biologic factors that are involved in the pathogenesis of NSCLC.	(24)
KIAA0101	PCNA-associated factor	Acts as a regulator of DNA repair during DNA replication. KIAA0101 expression in lung adenocarcinoma tissues is known to be higher than that in normal lung tissues according to a study.	(9)
KLC2	Kinesin light chain 2	Microtubule-associated force-producing protein that plays a role in organelle transport. KLC2 protein was found to be upregulated in NSCLC cell lines and tissues, and was an independent predictor of poor prognosis for elderly NSCLC patients.	(12)
MGAT4B	Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyl transferase B	Glycosyltransferase protein. MGAT4B were reported as oncogenic genes. It was also observed that MGAT4B transcripts were also upregulated in diethylnitrosamine-induced mouse model for hepatocellular carcinoma.	(8)
MLLT4	Afadin	Essential for the organization of adherens junctions. MLLT4, has been shown to be specific biomarkers for lung cancer epithelial cells in-situ.	(10)
PHLDA2	Pleckstrin homology-like domain family A member 2	Known to be tumor suppressor genes which is downregulated in lung cancer	(26)
PPARD	Peroxisome proliferator-activated receptor delta	They are known to function as a tumor suppressor, inhibiting development of primary tumors and metastases in lung cancer and other malignancies	(18)
PPP1R13L	RelA-associated inhibitor	PPP1R13L is prognostic, and high expression is unfavorable in lung cancer	(28)

PXDN	Peroxidasin homolog	May be a potential target for tumor immunotherapy, providing a new candidate that could improve cancer clinical diagnosis and treatment.	(29)
SCRIB	Protein scribble homolog	Low expression of SCRIB in CAFs is correlated with advanced tumor stages and poor survival for human lung squamous cell carcinoma.	(27)
SIN3B	Paired amphipathic helix protein Sin3b	Differentially regulates breast cancer	(14)
UBE2D4	Ubiquitin-conjugating enzyme E2 D4	UBE2T play critical roles in the progression of NSCLC and could be a potential therapeutic target for the treatment of NSCLC patients.	(15)