Mapping Relevant Genes in Severe Asthma: A Computational Strategy for Biomarker Discovery

Milenna Machado Pirovani[1,2][0000−0002−5060−9418], Leonardo Henrique da Silva[1][0000−0003−2341−5906], João Vítor Fernandes Dias[1][0000−0002−8156−9551], Luana Bastos[1], Letícia Gontijo[1], Rafael Lemos[1], Diego Mariano[1][0000−0002−5899−2052], Marcos Augusto dos Santos[1,3], and Raquel C. de Melo-Minardi[1,3][0000−0001−5190−100X]

[1] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[2] milennapirovani@hotmail.com
[3] Same contribution level.

Abstract. Asthma is a clinically heterogeneous, chronic inflammatory disorder of the airways, characterized by phenotypic variability that hampers accurate diagnosis and the development of tailored treatment strategies. The identification of robust molecular biomarkers is thus crucial for supporting disease stratification and informed therapeutic decision-making. In this study, we aimed to identify genes with high discriminatory power between patients with severe asthma and healthy controls by applying supervised statistical learning to publicly available transcriptomic data. We analyzed the gene expression dataset of severe asthmatic patients from the NCBI GEO repository using logistic regression for feature selection and classification. Our results suggest sets of genes potentially related to asthma, underscoring the importance of non-coding RNAs and pseudogenes. These results refine and complement previously published findings. Additionally, they highlight the value of reanalyzing public omics data with statistical models to generate novel hypotheses on disease mechanisms.

Keywords: Asthma · Biomarkers · Bioinformatics.

1 Introduction

Asthma is a chronic inflammatory disease of the airways, characterized by variable airflow obstruction, bronchial hyperresponsiveness, and persistent inflammation. These features are driven by the activation of immune cells — particularly eosinophils, Th2 lymphocytes, and airway epithelial cells — which promote the release of cytokines such as IL-4, IL-5, and IL-13 [12]. According to the World Health Organization, asthma affected approximately 262 million individuals worldwide in 2019, representing a major global health burden [7]. Multiple environmental and socioeconomic factors — including urbanization, exposure to aeroallergens, air pollution, respiratory infections, and health disparities — have contributed to the increasing prevalence and severity of asthma worldwide [25,31,11,3,2].

However, the lack of validated molecular biomarkers for clinical use hinders the accurate stratification of asthma endotypes, hindering the implementation of precision medicine-based therapeutic strategies [29,5]. Identifying these biomarkers is crucial for advancing personalized disease management and developing more effective targeted therapies. Studies in the literature have investigated various biomarkers associated with immune cell activation, airway remodeling, and inflammatory pathways. For example, IL-5, periostin, fractional exhaled nitric oxide (FeNO), blood eosinophil count, and serum IgE have been widely evaluated as indicators of inflammations predominant in many asthma phenotypes [8,27].

Transcriptomic analyses have played a crucial role in biomarker discovery. The study by [21] identified transcriptomic signatures associated with severe asthma from bronchial epithelial cells, revealing gene expression profiles linked to corticosteroid resistance and exacerbated inflammation. Similarly, [1] analyzed gene expression in sputum and reported that high levels of IL1RL1 and CLCA1 were associated with high Th2 endotypes. Furthermore, [5] highlighted the usefulness of multi-omics integration, combining transcriptomics, proteomics, and metabolomics, to refine biomarker selection and improve their clinical application. They utilized TPMS technology to analyze 94 candidate biomarkers for differentiating asthma and respiratory allergy phenotypes, and identified 16 molecular motifs and 12 proteins as being specifically associated with allergic asthma. Genes such as AKT1, STAT1, and MAPK13 were relevant to various diseases, while TLR4 was specific to asthmatic forms. This study enabled the prioritization of biomarkers based on their function in diseases, contributing to more accurate diagnoses [5].

Despite these advances, the validation and standardization of biomarkers across different populations and clinical settings still pose significant challenges. Therefore, the use of omics technologies, particularly high-resolution transcriptomic analysis, has proven to be a promising tool for identifying gene signatures associated with asthma severity. Public databases, such as the Gene Expression Omnibus (GEO) [4], provide datasets that enable the discovery of new biomarkers and a deeper understanding of the disease's biological structure.

In light of this, the combination of dimensionality reduction techniques, such as Singular Value Decomposition (SVD), with logistic regression models has demonstrated high effectiveness in biomedical scenarios involving large-scale gene expression data. For example, [10] utilized SVD to extract relevant components from human tumor microarray data, followed by the application of logistic regression to classify cancer types, resulting in significant gains in diagnostic accuracy. More recently, [22] applied SVD followed by penalized multinomial regression to classify medulloblastoma subtypes based on methylation data, reducing the number of probes from 320,000 to 200 components, which lead to an accuracy of roughly 99%. This evidence suggests that an integrated machine-learning approach may also be auspicious for complex respiratory diseases, such as asthma.

In this study, we applied regularized logistic regression to transcriptomic data from a publicly available gene expression dataset to identify biomarkers that can distinguish between patients with severe asthma and healthy controls. We also used a set of strategies to determine a list of genes with a potential link to the disease, ranging from information gain analysis to verification of the weight distribution (α) derived from the logistic regression model. We hope that the results presented here can complement and refine traditional microarray data analyses, which typically focus on gene expression differentiation.

2 Material and methods

2.1 Data collection

We collected the gene expression dataset of severe asthmatics, mild asthmatics, and healthy controls (access number: GSE27011), which was obtained from the Gene Expression Omnibus (GEO) repository [26]. This dataset was generated from a microarray gene expression experiment using the Affymetrix Human Genome 1.0 ST Array platform (GPL6244).

The GSE27011 dataset comprises 28,231 rows, corresponding to gene transcription data, and 54 columns, corresponding to individual samples. Subsequently, the dataset was transposed, resulting in a matrix of size $n \times m$, where $n = 54$ and $m = 28,231$. Each individual was classified into one of three target categories: Healthy control ($n = 18$), Mild asthma ($n = 19$), and Severe asthma ($n = 17$).

2.2 Data preprocessing

After downloading, the data were extracted, processed, and converted into CSV (Comma-Separated Values) and TSV (Tab-Separated Values) formats. Additionally, a sub-dataset with 35 entries was produced, containing only individuals classified into the Healthy control ($n = 18$) and Severe asthma ($n = 17$) categories. From now on, the main dataset, which contains the three classes, and the sub-dataset, which contains only the classes "Healthy control" and "Severe asthma", will be referred to as DS1 ($n = 54$) and DS2 ($n = 35$), respectively.

2.3 Machine learning classification model

To verify whether the experimental data could be correctly classified, supervised machine learning models were built using the Orange Data Mining tool v3.38 [6]. We built Logistic Regression models for DS1 and DS2 using the "Lasso - L1" and "Ridge - L2" regularization types ($c = 1$). The models were trained using stratified cross-validation ($k = 5$). We also compared models built with three supervised learning algorithms: KNN, SVM, and Random Forest. Default parameters were used for these models. For the KNN model, we used the parameters $K = 3$, Euclidean metric, and distance weighting. For SVM, we used cost

$(c = 1)$, regression loss epsilon of 0.1, and a sigmoid kernel. Lastly, for Random Forest, we set the number of trees to 100.

Then, the input matrix was dimensionally reduced using the SVD technique, and these data were used to compare with previously produced models. Details about SVD implementation will be presented in the next section.

2.4   Singular Value Decomposition (SVD)

Dimensional reduction was performed using the Singular Value Decomposition (SVD) technique in the MATLAB environment. In this approach, the matrix A was subjected to singular value decomposition (SVD) [20], such that: $A = U \cdot S \cdot V^T$ where $T \in \mathbb{R}^{m \times m}$ contains the left singular vectors (related to the samples); $S \in \mathbb{R}^{m \times n}$ is a rectangular matrix containing the singular values along the main diagonal and zeros elsewhere; $V \in \mathbb{R}^{n \times n}$ contains the right singular vectors (related to the attributes). To speed up the matrix decomposition process, the MATLAB "econ" parameter was utilized. In this case, the generated matrices have sizes $U = m \times r$, $S = r \times r$, and $V = n \times r$, where $r = min(m, n)$.

2.5   Modeling via Modified Logistic Regression

The methodology adopted to determine the Modified Logistic Regression was adapted from the work of [23] . Following singular value decomposition, another logistic regression model was constructed using MATLAB, with the additional objective of determining the importance of genes based on the classification of samples based on their gene expression patterns. We aim to estimate a weight vector $\alpha \in \mathbb{R}^n$ that defines the relative contribution of each attribute (gene) to the sample classification.

In this step, the i-th row of the transposed matrix A is represented by a vector $x \in \mathbb{R}^p$, where p is the number of attributes (genes) considered. The probability $P(x)$ of a sample belonging to the positive class (*i.e.*, associated with the clinical condition of interest – in this case, severe asthma) is modeled by the logistic function:

$$P(x) = \frac{exp(g(x))}{1 + expexp(g(x))}, with \ g(x) = x^t a \qquad (1)$$

where x is the feature vector for a sample, a is the vector of model coefficients, and $g(x)$ is the linear combination of features weighted by their respective coefficients.

To estimate α, the following regularized linear system was used:

$$\begin{bmatrix} I_m & -A \\ -A & I_n \end{bmatrix} \begin{bmatrix} x \\ \alpha \end{bmatrix} = \begin{bmatrix} -b \\ 0 \end{bmatrix} \qquad (2)$$

where $A \in \mathbb{R}^{m \times n}$ is the transposed gene expression matrix (patients $\times$ genes); I_m and I_n are identity matrices of dimensions m and n, respectively;

and $b \in \mathbb{R}^m$ is a vector encoding the log-odds function. For positive class samples: $b_i = \log(0.999/0.001)$. For negative class samples: $b_i = \log(0.001/0.999)$. The estimated probabilities were visualized to assess class separation based on the trained model (Supplementary Figure S3).

2.6   Attribute selection and construction of the reduced model

To reduce dimensionality and improve model interpretability, attribute selection was performed based on the magnitude of the coefficients in the α vector [19,24]. The coefficients were sorted in ascending order, and the 10 most negative and 10 most positive values were selected. Subsequently, the gene expression matrix A was reduced to a new matrix $A_r \in \mathbb{R}^{m \times 20}$, containing only the selected attributes. From A_r, a new weight vector α_r was recalculated by solving the regularized system again. The sample probabilities based on the selected attributes were then estimated according to the previously described logistic model.

3   Results and discussion

3.1   Classification models

In this study, we collected the gene expression dataset for asthmatics (GSE27011) from the GEO repository (here called DS1). This dataset consisted of 28,231 rows containing gene transcription information and 54 columns corresponding to individuals with severe asthma, mild asthmatics, and healthy controls. To better understand the aspects that differentiate severe asthma cases from the control group, a sub-dataset composed only of these two classes was created (here called DS2). This filtering strategy aimed to improve group separability by focusing on the immunologically distinct severe asthma phenotype [21]. Thus, both datasets were transposed, meaning the rows now represent individuals and the columns represent gene transcription data.

Initially, we verified whether we could classify the data based on the attributes we had collected. To do this, the datasets were imported into the Orange Data Mining tool, and logistic regression models were built using Lasso (L1) and Ridge (L2) regularization types. Table 1 presents the results of the constructed models.

For DS1, the classification using Lasso L1 logistic regression achieved an accuracy of 87%, while for Ridge L2, the accuracy was ~89% (Table 1, lines 1-2). For DS2, the accuracy was ~91% for L1 and ~97% for L2 (Table 1, lines 3-4). This demonstrates that the "mild asthma" class complicates model separation, since the transcription data can contain characteristics of both individuals without asthma (control group) and individuals with severe asthma. Removing this class, seen in DS2, improved the accuracy of both models. We can see this more clearly when considering other metrics, such as the harmonic mean (F1), which takes into account the precision and recall obtained in the results using "severe asthma" as the target class. In this case, the L1 model achieved an F1 score

Table 1. Results of logistic regression models built for datasets DS1 and DS2

SVD	Dataset	Logistic Regression	Accuracy	F1	Precision	Recall	Specificity
Without SVD	DS1	Lasso (L1)	0.870	0.788	0.812	0.765	0.919
		Ridge (L2)	0.889	0.824	0.824	0.824	0.919
	DS2	Lasso (L1)	0.914	0.914	0.889	0.941	0.889
		Ridge (L2)	0.971	0.970	1.000	0.941	1.000
With SVD (factor 8)	DS2	Lasso (L1)	0.971	0.971	0.944	1.000	0.944
		Ridge (L2)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

of 0.788 for DS1, compared to 0.914 for DS2, while the L2 model achieved an F1 score of 0.824 for DS1, compared to 0.970 for DS2 (Table 1, lines 1-4). The L2 model obtaining a better result than L1 means that in the dataset there are variables that contribute little, but make a difference in the final prediction.

Additionally, to verify whether the logistic regression model was the most appropriate for handling microarray data, we conducted similar experiments using three other machine learning algorithms: KNN, Random Forest, and SVM. The logistic regression models outperformed the other algorithms for classification in all datasets evaluated (Table 2).

Table 2. Comparison among Logistic Regression models (L1 and L2), Random Forest, SVM, and KNN models

Model	Accuracy			F-score		
	DS1	DS2	DS2+SVD	DS1	DS2	DS2+SVD
Logistic Regression (Ridge L2)	0.889	0.971	**1.000**	0.824	0.970	**1.000**
Logistic Regression (Lasso L1)	0.870	0.914	0.971	0.788	0.914	0.971
Random Forest	0.833	0.914	0.943	0.690	0.909	0.944
SVM	0.778	0.886	0.886	0.667	0.882	0.875
kNN	0.741	0.829	0.743	0.462	0.786	0.690

The partial conclusion of this section is that the proposed machine learning models performed well in classifying the data into the proposed classes, both for DS1 and DS2. Thus, we proceeded with other strategies to determine which genes contribute most to this classification.

3.2   Noise reduction using SVD

A significant challenge observed in this dataset is the immense number of attributes for each individual (each entry has over 28,000 features corresponding to gene expression levels). In machine learning problems, vast amounts of data for each instance can lead to errors in classification models, as they can introduce noise into the data. Therefore, we apply the singular value decomposition (SVD) technique to reduce the dataset and thus remove noise.

The SVD technique enables us to detect the singular values of the matrix used in the input, thereby reducing the search sample space and highlighting only the most relevant information for classification. In this case, we reduced the input matrix in DS2 ($n = 35$ and $m = 28,231$) to an auxiliary matrix with singular values ($n = 35$ and $m = 35$). SVD decomposes the original matrix into three matrices: U, S, and V, where matrix S stores the singular values. Plotting the diagonal of the "S" matrix shows the number of dimensions most relevant to representing the dataset. Dimensions (x-axis) after the line's elbow are less relevant and may represent noise that hinders data classification (and can therefore be removed). The factor plot indicated that only three dimensions of the auxiliary matrix could explain ∼80% of the variance (Supplementary Figure S1).

Additionally, we verified the importance of adding each dimension of the auxiliary matrix to the classification models. The accuracy of the L1 and L2 models for DS2 is illustrated in subsection 3.2. In this case, 35 models were built, incrementing a dimension with each new model for training. For example, the first model uses only the first dimension (the most important component), the second model uses the first and second dimensions, the third model uses dimensions 1, 2, and 3, and so on.

subsection 3.2 also shows that when eight dimensions (factor 8) were used to train the L2 model, the accuracy achieved reached 100% (details of this result can be found in the last two lines of Table 1). Furthermore, with only four dimensions, the models achieve an accuracy superior to 90%. After including five dimensions in the model, the accuracy remains high, ranging from ∼90% to 100%.
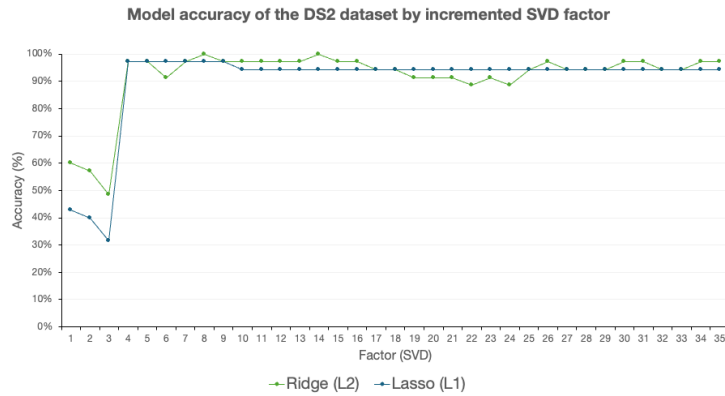

Fig. 1. Accuracy of the model when each dimension resulting from the reduced matrix was introduced as input to the model.

3.3   Most important genes

In the study by [26], the authors identified 1,378 genes differently expressed between individuals in the "Healthy control" and "Severe asthma" classes. Of these genes, they detected genes previously related to asthma in the literature, such as *INPP4A*, *PEX*, *RORA*, *PDE4D*, *IL2RB*, *TBX21*, *MYB*, *ORMDL3*, *PAFAH2*, *ADRB2*, *PTGDR*, and *HAVCR1*. However, gene expression difference analysis can return a large amount of data, which can be challenging to analyze manually. Therefore, we assessed the explainability from our models using the Information Gain algorithm. This algorithm returns the most important features for classifying the inputs, assigning each a score.

Table 3 presents the 10 most important features for classifying the models built for the DS2 dataset.

Table 3. Top 10 most important genes according to the Information Gain algorithm

#	Gene ID	Gene name	Description	Info. gain
1	7944525	*RNU6-1123P*	RNA, U6 small nuclear 1123, pseudogene	0.746
2	8005829	*PYY2*	peptide YY, 2 (pseudogene)	0.746
3	8157727	*GPR21*	G protein-coupled receptor 21	0.746
4	7907531	*GPR52*	G protein-coupled receptor 52	0.746
5	7968295	*RNU6-82P*	RNA, U6 small nuclear 82, pseudogene	0.746
6	8117018	*RNU6-522P*	RNA, U6 small nuclear 522, pseudogene	0.746
7	8144512	*RNU6-682P*	RNA, U6 small nuclear 682, pseudogene	0.746
8	7928489	*KAT6B*	K(lysine) acetyltransferase 6B	0.746
9	8103620	*RNU6-1336P*	RNA, U6 small nuclear 1336, pseudogene	0.673
10	8151411	*TCEB1*	Transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)	0.617

To confirm the differentiation between the gene expressions in the "Healthy control" and the "Severe asthma" classes pointed out by the Information Gain analysis, we present a heat map below (Table 3.3).

Table 3.3 confirms that the genes identified by the Information Gain algorithm (Table 3) have expression variations consistent with their assigned class. We can observe that a group of higher-expressed genes is present in the columns for "Severe asthma", while less expressed genes are found in the control group section (except for *RNU6-1336P*, where the opposite is true). This indicates that the genes described there may be related to asthma, but this affirmation further analysis.

Notably, Table 3 shows small nuclear and non-coding RNA transcription genes. The importance of non-coding RNAs has been increasingly studied recently. For instance, we can cite snRNAs, which participate in pre-mRNA splicing, a process that regulates the production of protein isoforms. Our results highlight that, in diseases such as asthma, particularly in severe cases, there are alterations in gene expression that may affect the immune and inflammatory
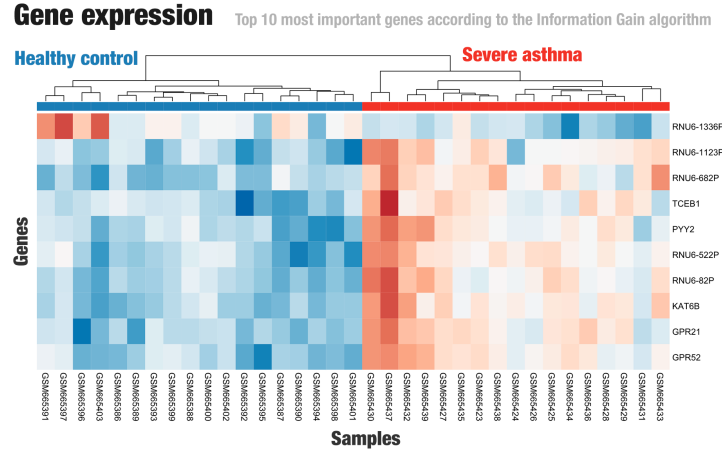
Fig. 2. Expression levels for the top 10 most important genes by the Information Gain algorithm. Values were normalized by row. Red cells: higher gene expression levels; blue cells: lower gene expression levels. Details are available at Supplementary Table S1.

response. Therefore, it makes sense that genes linked to snRNAs stand out as essential biomarkers.

With the results provided by the Information Gain algorithm, we sought a strategy to identify genes with potential to serve as biomarkers. Hence, we used the weight distribution (α) derived from the logistic regression model to determine the 20 genes that most positively and negatively impacted classification.

3.4   Feature selection using Logistic Regression

The weight distribution (α) derived from the logistic regression model indicates that most genes contribute minimally to classification, as evidenced by coefficients near zero (subsection 3.4A). A small subset exhibited high-magnitude weights, either positive or negative, suggesting their potential relevance for group discrimination. The ten genes with the highest and the ten genes with the lowest absolute α values were selected for downstream analysis (subsection 3.4B). Additionally, we can visually see the separation between the "health control" and "severe asthma" classes in a scatter plot of the principal components (subsection 3.4C).

Table 4 summarizes the main findings. Analysis of the proposed genes expression levels by the logistic regression model, show a clustering of genes that contribute positively (Supplementary Figure S2-A). However, we cannot see the same for genes that contribute negatively (Supplementary Figure S2-B).

Transcriptomic studies in severe asthma reveal differential expression not only of protein-coding genes, but also of non-coding RNAs, including antisense
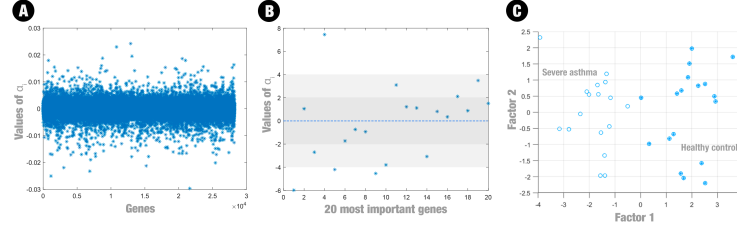
Fig. 3. Logistic regression results. **(A)** Projection of weight distribution (α) derived from the logistic regression model. **(B)** 20 genes that most positively (10) and negatively (10) impacted classification. **(C)** Two-dimensional representation of the projection of the 35 samples in the space. Healthy control: filled circles; severe asthma: empty circles.

Table 4. The top 20 genes that contribute most positively (10 for Severe asthma class) and negatively (10 for Healthy controls class) to the model according to the logistic regression model

	Positive			Negative		
#	**Gene ID**	**Gene name**	**Description**	**Gene ID**	**Gene name**	**Description**
1	8117018	*MYLIP*	Myosin regulatory light chain interacting protein	8137264	*TMEM176A*	Transmembrane protein 176A
2	7908861	*OCR1*	Ovarian cancer-related protein 1	7983890	*GRINL1A*	Myocardial zonula adherens protein
3	7967028	*RNU4-2*	RNA, U4 small nuclear 2	8110971	*CMBL*	Carboxymethylenebutenolidas homolog
4	7928489	*KAT6B*	Lysine acetyltransferase 6B	7904967	*RNVU1-19*	Variant U1 small nuclear 19
5	8052581	*ENSG00000278523*	Novel ncRNA chr:2:61928634-61928704	8106025	*BDP1*	Subunit of RNA polymerase III transcription initiation factor IIIB
6	8154211	*JAK2*	Janus kinase 2	7985431	*GOLGA2P3Y*	Golgin A2 pseudogene 3, Y-linked
7	7968295	*RNU6-82P*	RNA, U6 small nuclear 82, pseudogene	8031570	*RFPL4A*	Ret finger protein like 4A
8	8031152	*RPS9*	Ribosomal protein S9	8007848	*MAPK8IP1*	Mitogen-activated protein kinase 8 interacting protein 1 pseudogene
9	8055639	*ZEB2*	Zinc finger E-box binding homeobox 2	8029693	*FOSB*	FosB proto-oncogene
10	7975453	*SNORD56B*	Small nucleolar RNA, C/D box 56B	7914180	*SPCS2*	Signal peptidase complex subunit 2

transcripts, pseudogenes and lncRNAs [15]. These findings highlight the potential involvement of unconventional transcripts in disease mechanisms.

Among the genes identified, *JAK2* (Table 4) stands out due to its well-established for its central role in inflammatory cytokine signaling in the airways (IL-6, IL-4, IL-13) through activation of the JAK2/STAT3 pathway, targets of experimental inhibitors that have shown reductions in hyper-responsiveness and inflammation in both animal and human models [28]. The JAK2/STAT3 pathway influences asthma development by promoting inflammation and ferroptosis in airway epithelial cells. When cytokine IL-13 stimulates this pathway, it leads to the upregulation of *STAT3* and *JAK2* expression, which in turn increases *EPAS1* expression. This activation enhances ferroptosis (a form of iron-dependent cell death) and inflammatory responses, contributing to airway damage and asthma progression. Specifically, increased JAK2/STAT3 signaling exacerbates ferroptosis and inflammation by upregulating *EPAS1*, thereby amplifying airway epithelial injury associated with asthma [18].

Other genes in Table 4 exhibit more indirect or emerging associations with asthma. In a genome-wide association study (GWAS) identified an SNP (rs35514893) in the *DTNBP1-MYLIP* region associated with reduced risk of exacerbations under corticosteroid therapy ($OR \approx 0.36$, $p \approx 2.9 \times 19^{-6}$), suggesting possible modulation of steroid sensitivity, though *MYLIP* itself has not been directly implicated in asthma [13].

Regarding non-coding RNAs, *RNU4-2* which was upregulated in human bronchial epithelial cells (NHBE) after PGAP3 stimulation (log2 fold-change $\approx$ 7.01; $p = 0.008$), although its role in asthma is unclear [17]. Similarly, *SNORD56B* and *RNVU1-19* have no confirmed links to asthma, despite associations with systemic inflammation or weak CNV signals, respectively [30,9].

In contrast, *ZEB2*, a regulator of epithelial-mesenchymal transition (EMT), may influence airway remodeling. Diesel particle exposure upregulated ZEB2 in nasal epithelial cells, and its inhibition reduced polyp formation in murine models, implying a role in allergic airway inflammation [16].

Finally, FOSB, part of the AP-1 transcription factor family, was upregulated in bronchial epithelial cells under IL-17A stimulation in severe asthma. Its expression correlated with neutrophilic inflammation and disease severity, suggesting potential involvement in inflammatory signaling, although not yet defined as a susceptibility gene [14]. For the remaining genes, no clear associations with asthma have been documented to date.

3.5   Limitations

A major limitation of this work is the size of the datasets used ($n = 54$). In this study, we only used $k = 5$ cross-validation (scripts and datasets are available in the supplementary material repository). However, data obtained from microarray analyses present a special problem class, with a large number of columns ($m = 28,231$). In this context, dimensional reduction techniques, such as SVD, can represent a good solution for representing these data and removing noise. Furthermore, models based on logistic regression can handle this type of data

more effectively than other machine learning algorithms, as demonstrated in our analyses (Table 2). This corroborates other work available in the literature [23]. In the future, we intend to conduct additional studies to better evaluate the list of genes proposed here. Furthermore, experimental validations are necessary to confirm the results presented here.

4   Conclusion

This study aimed to validate previously reported asthma-associated genes and identify additional candidates with potential diagnostic or therapeutic relevance that have been little explored in the literature. Using logistic regression models, we propose a series of genes that may be related to asthma, but for which few studies are available. Our results highlight the importance of non-coding RNAs and pseudogenes. In future work, we plan to explore these results further, aiming to identify the role of these transcripts and the structural differences between genes and pseudogenes in severe asthma and healthy control groups. We hope that the results presented here can shed light on the mechanisms that lead to asthma and inform the development of new treatments.

References

1. Bigler, J., Boedigheimer, M., Schofield, J.P.R., Skipp, P.J., Corfield, J., Rowe, A., Sousa, A.R., Timour, M., Twehues, L., Hu, X., Roberts, G., Welcher, A.A., Yu, W., Lefaudeux, D., Meulder, B.D., Auffray, C., Chung, K.F., Adcock, I.M., Sterk, P.J., Djukanović, R.: A severe asthma disease signature from gene expression profiling of peripheral blood from u-biopred cohorts. Am J Respir Crit Care Med **195**(10), 1311–1320 (May 2017). `https://doi.org/10.1164/rccm.201604-0866oc`, `https://www.atsjournals.org/doi/10.1164/rccm.201604-0866OC`, publisher: American Thoracic Society
2. Chatkin, J., Correa, L., Santos, U.: External environmental pollution as a risk factor for asthma. Clinic Rev Allerg Immunol **62**(1), 72–89 (Feb 2022). `https://doi.org/10.1007/s12016-020-08830-5`, `https://link.springer.com/10.1007/s12016-020-08830-5`, publisher: Springer Science and Business Media LLC

3. Checkley, W.: What have we learnt from studies on the association between urbanisation and asthma prevalence? Thorax **74**(11), 1011–1012 (Nov 2019). `https://doi.org/10.1136/thoraxjnl-2019-213319`, `https://thorax.bmj.com/lookup/doi/10.1136/thoraxjnl-2019-213319`, publisher: BMJ

4. Clough, E., Barrett, T.: The gene expression omnibus database. In: Methods in Molecular Biology, pp. 93–110. Springer New York, New York, NY (2016). `https://doi.org/10.1007/978-1-4939-3578-9_5`, `http://link.springer.com/10.1007/978-1-4939-3578-9_5`, iSSN: 1064-3745, 1940-6029

5. Cremades-Jimeno, L., De Pedro, M.A., López-Ramos, M., Sastre, J., Mínguez, P., Fernández, I.M., Baos, S., Cárdaba, B.: Prioritizing molecular biomarkers in asthma and respiratory allergy using systems biology. Front. Immunol. **12** (Apr 2021). `https://doi.org/10.3389/fimmu.2021.640791`, `https://www.frontiersin.org/articles/10.3389/fimmu.2021.640791/full`, publisher: Frontiers Media SA

6. Demšar, J., Curk, T., Erjavec, A., Gorup, v., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: data mining toolbox in python. J. Mach. Learn. Res. **14**(1), 2349–2353 (Jan 2013)

7. Diseases, G.., Collaborators, I.: Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet **396**(10258), 1204–1222 (Oct 2020). `https://doi.org/10.1016/S0140-6736(20)30925-9`

8. Dweik, R.A., Boggs, P.B., Erzurum, S.C., Irvin, C.G., Leigh, M.W., Lundberg, J.O., Olin, A.C., Plummer, A.L., Taylor, D.R., on behalf of the American Thoracic Society Committee on Interpretation of Exhaled Nitric Oxide Levels (Fe) for Clinical Applications: An official ats clinical practice guideline: Interpretation of exhaled nitric oxide levels (feno) for clinical applications. Am J Respir Crit Care Med **184**(5), 602–615 (Sep 2011). `https://doi.org/10.1164/rccm.9120-11st`, `https://www.atsjournals.org/doi/10.1164/rccm.9120-11ST`, publisher: American Thoracic Society

9. Fawcett, K.A., Demidov, G., Shrine, N., Paynton, M.L., Ossowski, S., Sayers, I., Wain, L.V., Hollox, E.J.: Exome-wide analysis of copy number variation shows association of the human leukocyte antigen region with asthma in uk biobank. BMC Med Genomics **15**(1) (Dec 2022). `https://doi.org/10.1186/s12920-022-01268-y`, `https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-022-01268-y`, publisher: Springer Science and Business Media LLC

10. Ghosh, D.: Singular value decomposition regression models for classification of tumors from microarray experiments. Pac Symp Biocomput pp. 18–29 (2002)

11. Goldin, J., Hashmi, M.F., Cataletto, M.E.: Asthma. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2025), `http://www.ncbi.nlm.nih.gov/books/NBK430901/`

12. Hamid, Q., Tulic', M.K., Liu, M.C., Moqbel, R.: Inflammatory cells in asthma: mechanisms and implications for therapy. J Allergy Clin Immunol **111**(1 Suppl), S5–S12; discussion S12–17 (Jan 2003). `https://doi.org/10.1067/mai.2003.22`

13. Hernandez-Pacheco, N., Farzan, N., Francis, B., Karimi, L., Repnik, K., Vijverberg, S.J., Soares, P., Schieck, M., Gorenjak, M., Forno, E., Eng, C., Oh, S.S., Pérez-Méndez, L., Berce, V., Tavendale, R., Samedy, L.A., Hunstman, S., Hu, D., Meade, K., Farber, H.J., Avila, P.C., Serebrisky, D., Thyne, S.M., Brigino-Buenaventura, E., Rodriguez-Cintron, W., Sen, S., Kumar, R., Lenoir, M., Rodriguez-Santana,

J.R., Celedón, J.C., Mukhopadhyay, S., Potočnik, U., Pirmohamed, M., Verhamme, K.M., Kabesch, M., Palmer, C.N.A., Hawcutt, D.B., Flores, C., Maitland-van Der Zee, A.H., Burchard, E.G., Pino-Yanes, M.: Genome-wide association study of inhaled corticosteroid response in admixed children with asthma. Clin Experimental Allergy **49**(6), 789–798 (Jun 2019). `https://doi.org/10.1111/cea.13354`, `https://onlinelibrary.wiley.com/doi/10.1111/cea.13354`, publisher: Wiley

14. Jakiela, B., Soja, J., Sladek, K., Przybyszowski, M., Plutecka, H., Gielicz, A., Licholai, S., Aab, A., Rebane, A., Bochenek, G.: Bronchial epithelial cell transcriptome shows endotype heterogeneity of asthma in patients with nsaid-exacerbated respiratory disease. J Allergy Clin Immunol **151**(4), 953–965 (Apr 2023). `https://doi.org/10.1016/j.jaci.2022.10.029`

15. Jurkiewicz, M.M., Mueller-Alcazar, A., Moser, D.A., Jayatilaka, I., Mikhailik, A., Ferri, J., Fogelman, N., Canli, T.: Integrated microrna and mrna gene expression in peripheral blood mononuclear cells in response to acute psychosocial stress: a repeated-measures within-subject pilot study. BMC Res Notes **14**(1), 222 (Jun 2021). `https://doi.org/10.1186/s13104-021-05635-3`

16. Lee, M., Lim, S., Kim, Y.S., Khalmuratova, R., Shin, S.H., Kim, I., Kim, H.J., Kim, D.Y., Rhee, C.S., Park, J.W., Shin, H.W.: Dep-induced zeb2 promotes nasal polyp formation via epithelial-to-mesenchymal transition. Journal of Allergy and Clinical Immunology **149**(1), 340–357 (Jan 2022). `https://doi.org/10.1016/j.jaci.2021.04.024`, `https://linkinghub.elsevier.com/retrieve/pii/S0091674921007193`, publisher: Elsevier BV

17. Leslie, E., Miller, M., Lafuze, A., Svyatskaya, S., Choi, G.S., Broide, D.H.: Pgap3 regulates human bronchial epithelial cell mrnas present in asthma and respiratory virus reference data sets. medRxiv p. 2024.07.03.24309917 (Jul 2024). `https://doi.org/10.1101/2024.07.03.24309917`

18. Liu, L., Yang, C., Li, Y., Zhou, H., Shi, M., Shi, T., Shi, W.: Epas1 amplifies asthma pathogenesis through jak2/stat3-mediated ferroptosis and inflammation. Biomol Biomed (Mar 2025). `https://doi.org/10.17305/bb.2025.11334`, `https://www.bjbms.org/ojs/index.php/bjbms/article/view/11334`, publisher: Association of Basic Medical Sciences of FBIH

19. Liu, X.Y., Wu, S.B., Zeng, W.Q., Yuan, Z.J., Xu, H.B.: Logsum + l2 penalized logistic regression model for biomarker selection and cancer classification. Sci Rep **10**(1), 22125 (Dec 2020). `https://doi.org/10.1038/s41598-020-79028-0`

20. Marcolino, L.S., Couto, B.R.G.M., Dos Santos, M.A.: Genome visualization in space. In: Advances in Intelligent and Soft Computing, pp. 225–232. Springer Berlin Heidelberg, Berlin, Heidelberg (2010). `https://doi.org/10.1007/978-3-642-13214-8_29`, `http://link.springer.com/10.1007/978-3-642-13214-8_29`, iSSN: 1867-5662, 1867-5670

21. Modena, B.D., Bleecker, E.R., Busse, W.W., Erzurum, S.C., Gaston, B.M., Jarjour, N.N., Meyers, D.A., Milosevic, J., Tedrow, J.R., Wu, W., Kaminski, N., Wenzel, S.E.: Gene expression correlated with severe asthma characteristics reveals heterogeneous mechanisms of severe disease. Am J Respir Crit Care Med **195**(11), 1449–1463 (Jun 2017). `https://doi.org/10.1164/rccm.201607-1407OC`

22. Mohammed, I., Elbashir, M.K., Faggad, A.S.: Singular value decomposition-based penalized multinomial regression for classifying imbalanced medulloblastoma subgroups using methylation data. J Comput Biol **31**(5), 458–471 (May 2024). `https://doi.org/10.1089/cmb.2023.0198`

23. Morais-Rodrigues, F., Silvério-Machado, R., Kato, R.B., Rodrigues, D.L.N., Valdez-Baez, J., Fonseca, V., San, E.J., Gomes, L.G.R., Dos Santos, R.G., Vinicius

Canário Viana, M., da Cruz Ferraz Dutra, J., Teixeira Dornelles Parise, M., Parise, D., Campos, F.F., de Souza, S.J., Ortega, J.M., Barh, D., Ghosh, P., Azevedo, V.A.C., Dos Santos, M.A.: Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. Gene **726**, 144168 (Feb 2020). `https://doi.org/10.1016/j.gene.2019.144168`

24. Mori, Y., Yokota, H., Hoshino, I., Iwatate, Y., Wakamatsu, K., Uno, T., Suyari, H.: Deep learning-based gene selection in comprehensive gene analysis in pancreatic cancer. Sci Rep **11**(1), 16521 (Aug 2021). `https://doi.org/10.1038/s41598-021-95969-6`

25. Ni, R., Su, H., Burnett, R.T., Guo, Y., Cheng, Y.: Long-term exposure to pm2.5 has significant adverse effects on childhood and adult asthma: A global meta-analysis and health impact assessment. One Earth **7**(11), 1953–1969 (Nov 2024). `https://doi.org/10.1016/j.oneear.2024.09.022`, `https://linkinghub.elsevier.com/retrieve/pii/S2590332224004871`, publisher: Elsevier BV

26. Orsmark-Pietras, C., James, A., Konradsen, J.R., Nordlund, B., Söderhäll, C., Pulkkinen, V., Pedroletti, C., Daham, K., Kupczyk, M., Dahlén, B., Kere, J., Dahlén, S.E., Hedlin, G., Melén, E.: Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. Eur Respir J **42**(1), 65–78 (Jul 2013). `https://doi.org/10.1183/09031936.00077712`, `https://publications.ersnet.org/lookup/doi/10.1183/09031936.00077712`, publisher: European Respiratory Society (ERS)

27. Price, D.B., Rigazio, A., Campbell, J.D., Bleecker, E.R., Corrigan, C.J., Thomas, M., Wenzel, S.E., Wilson, A.M., Small, M.B., Gopalan, G., Ashton, V.L., Burden, A., Hillyer, E.V., Kerkhof, M., Pavord, I.D.: Blood eosinophil count and prospective annual asthma disease burden: a uk cohort study. Lancet Respir Med **3**(11), 849–858 (Nov 2015). `https://doi.org/10.1016/S2213-2600(15)00367-7`

28. Ren, Y., Liu, Y., Wang, S., Lei, Z., Yan, Y., Guan, X., Hou, J., Zhu, S., Shan, H., Tian, X., Wang, Q., Cao, C., Zhang, Y., Ma, Y.: Zhike pingchuan granules improve bronchial asthma by regulating the il-6/jak2/stat3 pathway. Exp Ther Med **22**(2), 899 (Aug 2021). `https://doi.org/10.3892/etm.2021.10331`

29. Skloot, G.S.: Asthma phenotypes and endotypes: a personalized approach to treatment. Curr Opin Pulm Med **22**(1), 3–9 (Jan 2016). `https://doi.org/10.1097/MCP.0000000000000225`

30. Vishnyakova, P., Kuznetsova, M., Poltavets, A., Fomina, M., Kiseleva, V., Muminova, K., Potapova, A., Khodzhaeva, Z., Pyregov, A., Trofimov, D., Elchaninov, A., Sukhikh, G., Fatkhudinov, T.: Distinct gene expression patterns for cd14++ and cd16++ monocytes in preeclampsia. Sci Rep **12**(1) (Sep 2022). `https://doi.org/10.1038/s41598-022-19847-5`, `https://www.nature.com/articles/s41598-022-19847-5`, publisher: Springer Science and Business Media LLC

31. Zheng, X.Y., Xu, Y.J., Guan, W.J., Lin, L.F.: Regional, age and respiratory-secretion-specific prevalence of respiratory viruses associated with asthma exacerbation: a literature review. Arch Virol **163**(4), 845–853 (Apr 2018). `https://doi.org/10.1007/s00705-017-3700-y`