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

**Milenna Machado Pirovani 1,\*, [...]Leonardo Henrique da Silva1, Marcos Augusto dos Santos 1,†  and Raquel C. de Melo-Minardi 1,†**

1 Laboratory of Bioinformatics and Systems (LBS), Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

2D…

milennapirovani@hotmail.com

**\***Correspondence author; **†**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 GSE27011 gene expression dataset from the NCBI GEO repository using L2-regularized logistic regression for feature selection and classification. These results underscore the utility of mining public transcriptomic datasets through machine learning to uncover underexplored molecular signatures of complex diseases. These findings 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 [[Hamid et al. 2003]](https://www.zotero.org/google-docs/?RADDrF). According to the World Health Organization, asthma affected approximately 262 million individuals worldwide in 2019, representing a major global health burden [[Vos et al. 2020]](https://www.zotero.org/google-docs/?ptBMHd). 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 [[Ni et al. 2024; Pearce et al. 2000; Zheng et al. 2018; Goldin et al. 2025; Checkley 2019; Chatkin et al. 2022]](https://www.zotero.org/google-docs/?wnkaHt).

Asthma represents a clinically heterogeneous disease, encompassing distinct phenotypes and endotypes that, despite sharing respiratory symptoms such as cough, dyspnea, and wheezing, present different underlying immunopathological mechanisms, leading to variable responses to treatment and significant prognostic differences. For example, patients with type 2 eosinophilic inflammation respond better to corticosteroids and targeted biologic therapies, while other subgroups, such as those with neutrophilic inflammation or those lacking classic inflammatory biomarkers, exhibit greater resistance to conventional treatments [[Agache and Akdis 2019; Wenzel 2012]](https://www.zotero.org/google-docs/?fM8Kqw).

In particular, 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 [[Skloot 2016; Cremades-Jimeno et al. 2021]](https://www.zotero.org/google-docs/?0E0f7g). 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 type 2 inflammation, which is predominant in many asthma phenotypes [[Dweik et al. 2011; Price et al. 2015]](https://www.zotero.org/google-docs/?0jP7E8) .

Transcriptomic analyses have played a crucial role in biomarker discovery. The study by Modena et al. (2017) [[Modena et al. 2017]](https://www.zotero.org/google-docs/?vdVRfG) identified transcriptomic signatures associated with severe asthma from bronchial epithelial cells, revealing gene expression profiles linked to corticosteroid resistance and exacerbated inflammation. Similarly, Bigler et al. (2017) [[Bigler et al. 2017]](https://www.zotero.org/google-docs/?oiwI6K) analyzed gene expression in sputum and reported that high levels of IL1RL1 and CLCA1 were associated with high Th2 endotypes. Furthermore, Cremades-Jimeno et al. (2021) [[Cremades-Jimeno et al. 2021]](https://www.zotero.org/google-docs/?RG3lCp) highlighted the usefulness of multi-omics integration, combining transcriptomics, proteomics, and metabolomics, to refine biomarker selection and improve their clinical application. This study utilized TPMS technology to analyze 94 candidate biomarkers for differentiating asthma and respiratory allergy phenotypes. The authors identified 16 molecular motifs associated with allergic asthma, highlighting 17 proteins related to this condition, 11 proteins related to respiratory allergy, and 16 proteins related to non-allergic asthma. Of these, 12 proteins were identified as being specifically associated with allergic asthma. Genes such as AKT1, STAT1, and MAPK13 were relevant to all three 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 [[Cremades-Jimeno et al. 2021]](https://www.zotero.org/google-docs/?FQSi6P).

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), 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, [[Ghosh 2002]](https://www.zotero.org/google-docs/?PDVUfw) 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, [[Mohammed et al. 2024]](https://www.zotero.org/google-docs/?za6ANT) applied SVD followed by penalized multinomial regression to classify medulloblastoma subtypes based on methylation data, reducing the number of probes from more than 320,000 to approximately 200 components and achieving a cross-validation accuracy of roughly 99%. This evidence suggests that the integrated approach may also be auspicious for complex respiratory diseases, such as asthma, enabling the identification of specific molecular signatures associated with clinical phenotypes and therapeutic responses.

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 severe asthma patients and healthy controls.

# 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 [[Orsmark-Pietras et al. 2013a]](https://www.zotero.org/google-docs/?1nHJMm). 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 x 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.

Subsequently, the datasets were imported by the Orange Data Mining tool and the MATLAB® environment [[Hamid et al. 2003]](https://www.zotero.org/google-docs/?3ifOiW). The resulting matrix consists of *n* columns corresponding to genes (attributes) and *m* rows representing samples (patients). For modeling purposes, the transpose of this matrix was considered, denoted as **A** ∈ ℝ ᵐ ˣ ⁿ, where each row of **A** represents a sample as a vector in ℝⁿ. The source code developed for this work is available in the supplementary materials.

## 2.3 Machine learning classification model

To verify whether the experimental data could be classified using machine learning, supervised machine learning models were built using the Orange Data Mining tool v3.38 [[Demšar et al. 2013]](https://www.zotero.org/google-docs/?CvSI4N).

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, metric Euclidean, and weighting by distance. For SVM, we used cost (c=1), regression loss epsilon of 0.1, and a sigmoid kernel. For Random Forest, the number of trees was set 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) [[Abdelwahab et al. 2023; Ni et al. 2024]](https://www.zotero.org/google-docs/?2iiB67), such that:

A = U·S·Vᵀ, (1)

where **T** ∈ ℝᵐˣᵐ contains the left singular vectors (related to the samples); **S** ∈ ℝᵐˣⁿ is a rectangular matrix containing the singular values along the main diagonal and zeros elsewhere; **V** ∈ ℝⁿˣⁿ 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* x *r*, S = *r* x *r*, and V = *n* x *r*, where *r* = min(*m*, *n*).

The distribution of singular values was plotted to analyze the variability explained by each component. Subsequently, the projections of the samples onto the first principal components were used for three-dimensional visualization of the entity space, allowing for a preliminary observation of class separation.

## 2.5 Modeling via Modified Logistic Regression

The methodology adopted to determine the Modified Logistic Regression was adapted from the work of Morais-Rodrigues et al. (<https://www.sciencedirect.com/science/article/pii/S0378111919308273>). 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 [[Genç 2024; Liu et al. 2020]](https://www.zotero.org/google-docs/?WXaHak). The objective was to estimate a weight vector α ∈ ℝⁿ 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 ∈ ℝP, 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) is modeled by the logistic function:

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:

where A ∈ ℝᵐˣⁿ is the transposed gene expression matrix (patients × genes); Iₘ and Iₙ are identity matrices of dimensions *m* and *n*, respectively; and b ∈ ℝᵐ is a vector encoding the log-odds function. For positive class samples: bᵢ = log(0.999/0.001). For negative class samples: bᵢ = log(0.001/0.999). The estimated probabilities were visualized to assess class separation based on the trained model.

## 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 [[Liu et al. 2020; Mori et al. 2021]](https://www.zotero.org/google-docs/?Drv0sN). The coefficients were sorted in ascending order, and the 10 most negative and 10 most positive values were selected, totaling 20 genes. Subsequently, the gene expression matrix A was reduced to a new matrix A∈ ℝmx20, containing only the selected attributes. From Aᵣ, a new weight vector αᵣ 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 [[Modena et al. 2017]](https://www.zotero.org/google-docs/?xN0xJK). 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 collected attributes. 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.

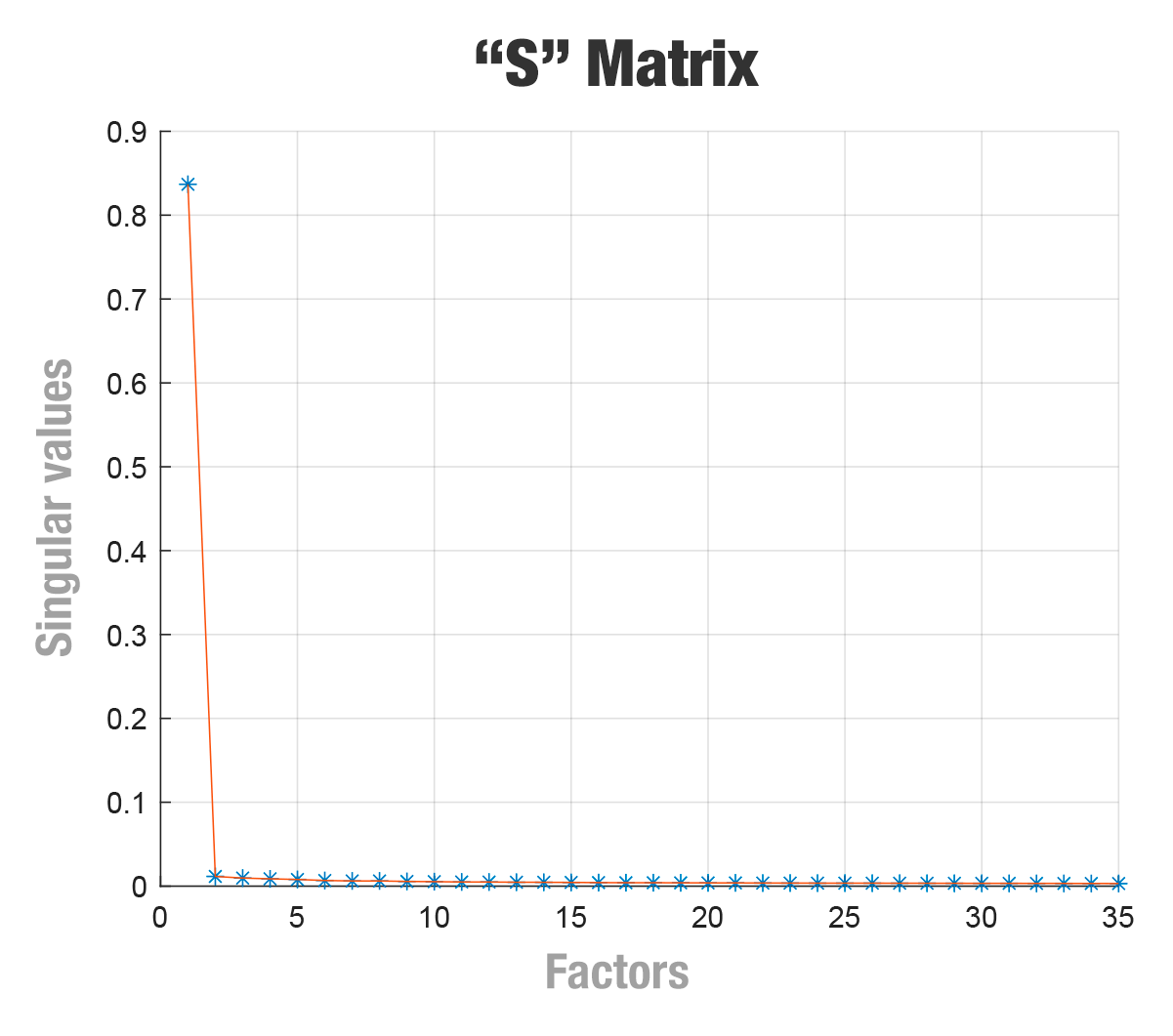
**Table 1. Results of logistic regression models built for datasets DS1 and DS2**

| **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** |

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 and 2). For DS2, the accuracy was ~91% for L1 and ~97% for L2 (Table 1, lines 3 and 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 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 to 4).

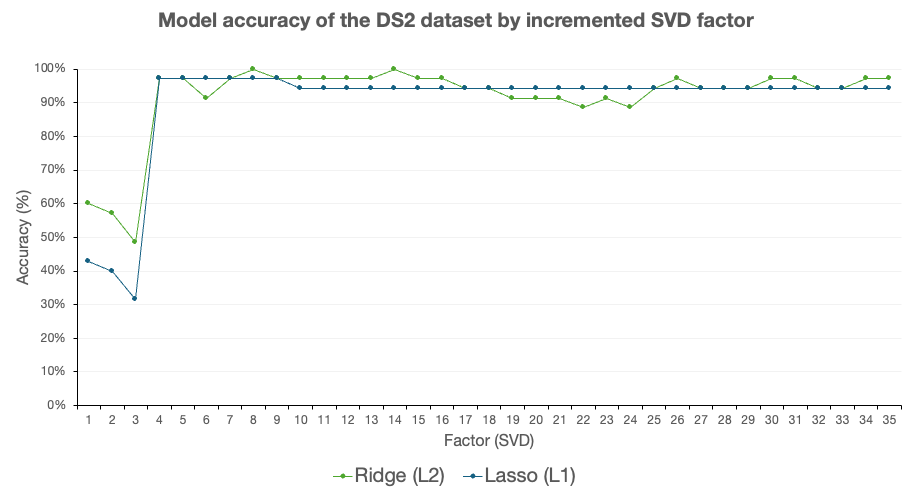
## 3.2 Noise reduction using SVD

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 ADS2 (n=54 and m=28,231) to an auxiliary matrix with singular values (n=35 and m=35). The factor plot indicated that only three dimensions of the auxiliary matrix could explain approximately 80% of the variance (Figure 1). 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).



**Figure 1. Importance of singular values by total factors (principal components).**

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 Figure 2. In this case, 36 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, the fourth model uses dimensions 1, 2, 3, and 4, and so on.



**Figure 2. Accuracy of the model when each dimension resulting from the reduced matrix was introduced as input to the model.**

Figure 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%.

## 3.3 Comparison to other machine learning algorithms

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 |

## 3.4 Most important genes

In the study by [[Orsmark-Pietras et al. 2013b]](https://www.zotero.org/google-docs/?gFcWiF), 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* (8043791), *PEX* (7921571), *RORA* (7989365), *PDE4D* (8112220), *IL2RB* (8075886), *TBX21* (8008029), *MYB* (8122202), *ORMDL3* (8014916), *PAFAH2* (7913883), *ADRB2* (8109086), *PTGDR* (7974363), and *HAVCR1* (8115455).

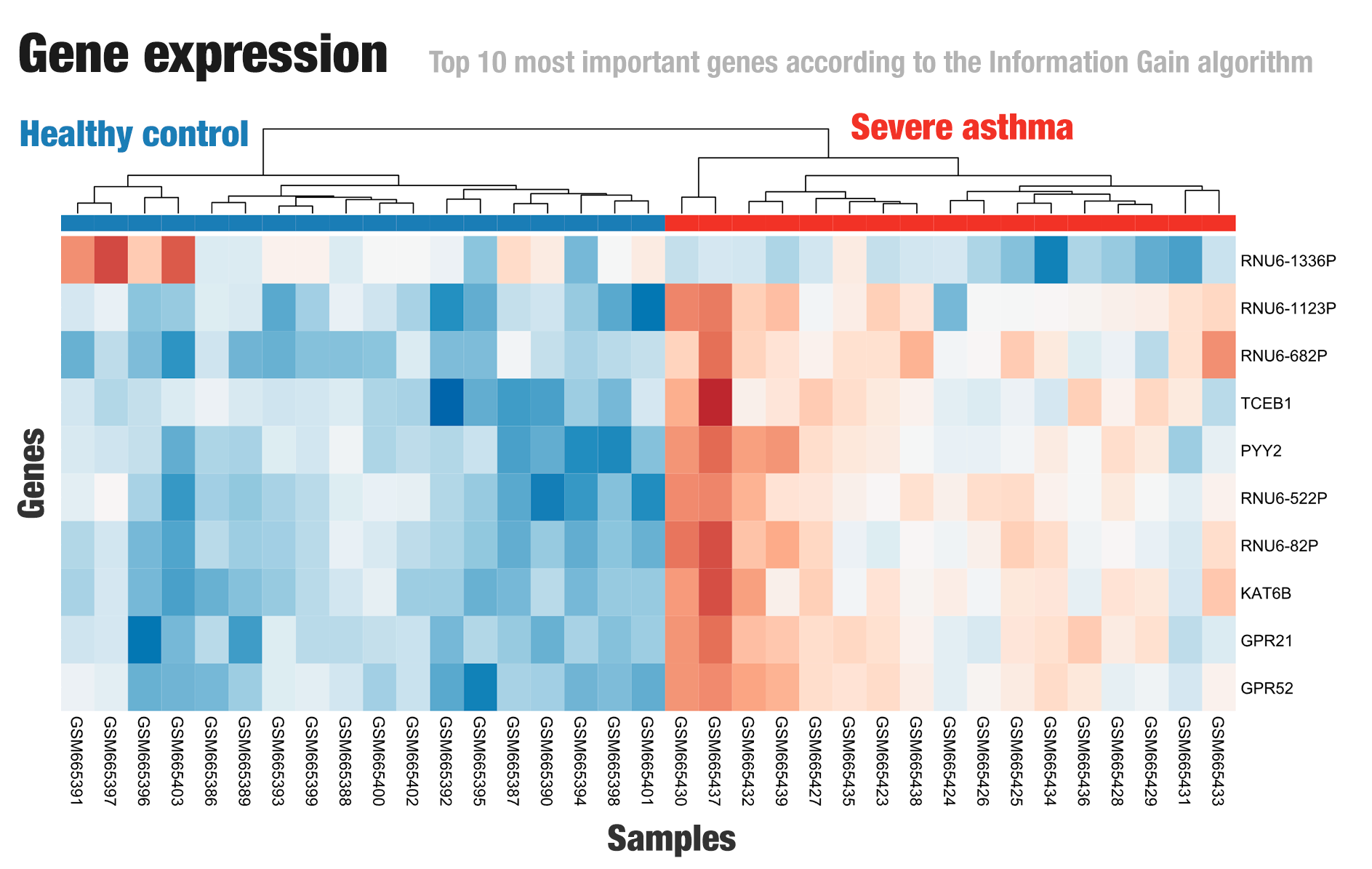
However, gene expression difference analysis can return a large amount of data, which can be challenging to analyze manually. Therefore, we assessed the explainability from models using the Information Gain algorithm. This algorithm returns the most important features for classifying the inputs, assigning 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 by 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 in the red-white-blue color scheme (Figure 3). In this map, red cells indicate higher gene expression levels, while blue cells indicate lower levels. Importantly, the color scale has been row-normalized to highlight differences between groups (see Supplemental Table S1 for details).



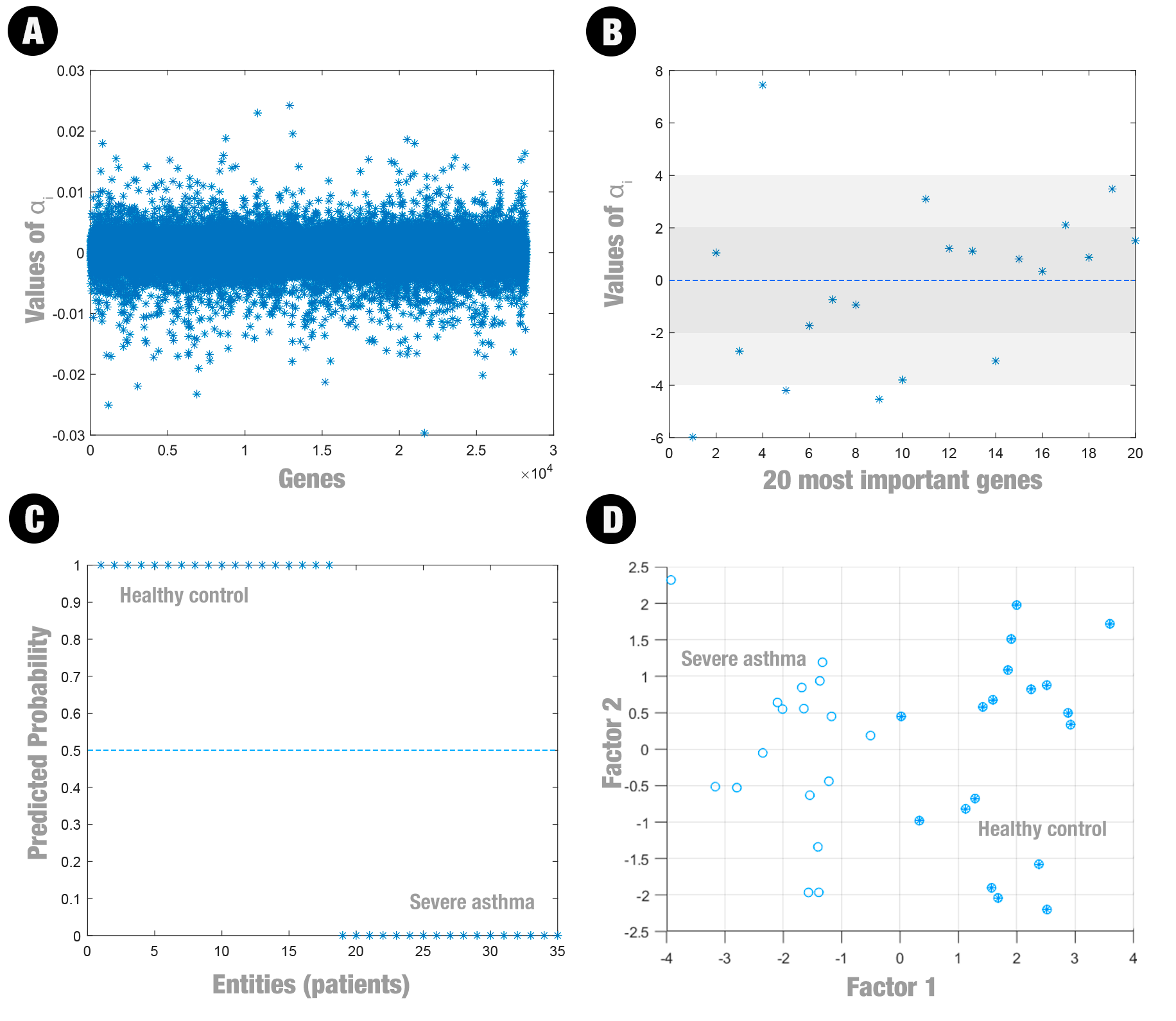
**Figure 3. Expression levels for the top 10 most important genes by the Information Gain algorithm. Values were normalized by row. Red cells: higher gene expression levels; blue cells: lower gene expression levels. Details are available at Supplementary Table S1.**

Figure 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. However, assessing the reason for this requires further analysis.

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.

## 3.5 Feature selection using Logistic Regression

The weight distribution (α) derived from the logistic regression model indicated that most genes contributed minimally to classification, as evidenced by coefficients near zero (Figure 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 10 genes with the lowest absolute α values were selected for downstream analysis (Figure 4B).



**Figure 4. (A) Projection of the Samples in Space of weight distribution (α) derived from the logistic regression model. (B) 20 genes that most positively (10) and negatively (10) impacted classification. (C) Classification Probabilities Obtained by Logistic Regression Using All Genes. Scores close to 1: control group; scores close to 0: severe asthma group. (D) Two-dimensional representation of the projection of the 35 samples in the space. The control group samples are highlighted with filled circles, while the severe asthma samples are represented by empty circles.**

Table 4 summarizes the main findings of this analysis.

**Table 4. The top 10 genes that contribute most positively (Severe asthma class) and negatively (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* | Carboxymethylenebutenolidase 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 |

Transcriptomic studies in severe asthma have shown that differential gene expression includes not only protein-coding genes, but also various categories of non-coding RNAs and other unconventional transcripts (<https://pmc.ncbi.nlm.nih.gov/articles/PMC8176593/#:~:text=RNU6,0> ). For example, patients with severe asthma exhibit widespread alterations in the expression of non-coding transcripts (antisense RNAs, pseudogenes, and IncRNAs) (<https://pmc.ncbi.nlm.nih.gov/articles/PMC8176593/#:~:text=RNU6,0> ), illustrating that such molecules may appear in gene expression profiles.

It’s interesting to note the association with *JAK2* (table ? - feature selection), already recognized 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 (<https://pmc.ncbi.nlm.nih.gov/articles/PMC8243345/?utm_source=chatgpt.com> ). 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 (<https://www.researchgate.net/publication/390262712_EPAS1_amplifies_asthma_pathogenesis_through_JAK2STAT3-mediated_ferroptosis_and_inflammation> ).

The other identified genes (table 4) have indirect evidence or potential associations with asthma. In a genome-wide association study (GWAS) focused on response to inhaled corticosteroids in asthmatic children, a polymorphism (SNP rs35514893) in the DTNBP1-MYLIP region showed a significant association with reduced exacerbations under treatment (OR≈0.36, p≈2.9 × 10⁻⁶). This suggests that variants linked to MYLIP might influence corticosteroid-mediated asthmatic inflammation, although there is no evidence that MYLIP itself is a primary risk gene for asthma ( <https://pmc.ncbi.nlm.nih.gov/articles/PMC7054824/#:~:text=%280,046> ). (LINK) An *in vitro* study using human bronchial epithelium (NHBE cells) found that *RNU4-2* expression was strongly induced by the PGAP3 protein (log2 fold-change ≈7.01; p=0.008). Although this reflects inflammatory transcriptional regulation in respiratory cells, there are no clinical studies directly linking *RNU4-2* to asthma. *RNU4-2* is a small nuclear RNA (U4 snRNA), and its role in the respiratory immune response is not yet fully understood (<https://pmc.ncbi.nlm.nih.gov/articles/PMC11245055/#:~:text=7%20%282%20%29%20BCO1%20%28Beta,0.001> ). The SNORD56B is a small nucleolar RNA with no known associations with asthma. It was found among the most dysregulated RNAs in monocytes from patients with preeclampsia (underexpression of SNORD56B), indicating a possible involvement in systemic inflammatory processes. However, there is no evidence of SNORD56B involvement in allergic or asthmatic pathways (<https://pmc.ncbi.nlm.nih.gov/articles/PMC9474473/#:~:text=oxidized%20low,and%20down> ). Such as the RNU1-19, which is a variant with no known association with asthma. A copy number variation (CNV) analysis briefly reported a region involving RNU1-19 with a modest association (p≈0.012) in asthma, but this finding was not confirmed and appears statistically weak. Therefore, RNU1-19 is not currently considered a gene linked to asthma (<https://pmc.ncbi.nlm.nih.gov/articles/PMC9124406/#:~:text=...%20pmc.ncbi.nlm.nih.gov%20%20NBPF26%2FPPIAL4A%2FRNVU1,2%5D%20have%20been> ).

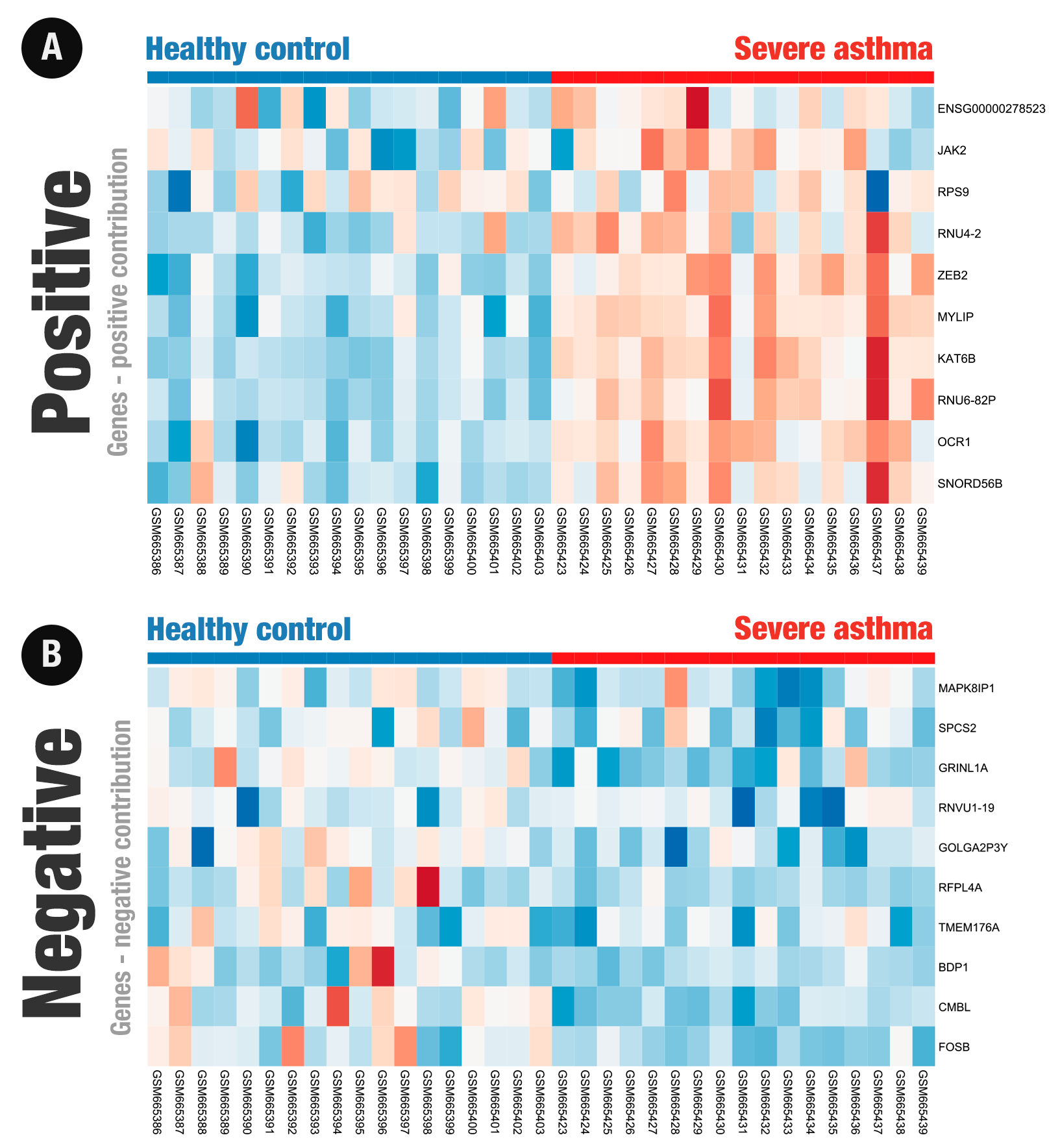
(LINK) ZEB2is a transcription factor that regulates epithelial-mesenchymal transition (EMT), a process involved in airway remodeling. Lee et al. demonstrated that diesel particles induce EMT in human nasal epithelial cells via ZEB2, and that inhibiting ZEB2 prevent nasal polyp formation in mice. Although this study focused on nasal polyps (an allergic respiratory disease), it suggests that ZEB2 may mediate inflammatory responses and remodeling associated with asthma and allergic respiratory conditions ( <https://www.jacionline.org/article/S0091-6749(21)00719-3/fulltext> ).

(LINK) *FOSB* encodes a member of the AP-1 family of transcription factors. Transcriptome studies of bronchial epithelium in severe asthma have identified a proinflammatory profile induced by IL-17A, in which *FOSB* showed elevated expression. This proinflammatory profile was correlated with greater clinical severity and neutrophilia, and *FOSB* expression was associated with asthma severity. Therefore, *FOSB* may influence inflammatory signaling in asthma, although its direct genetic role in disease susceptibility remains undefined (<https://www.jacionline.org/article/S0091-6749(22)01511-1/fulltext> ).

For the others, there are no published studies associating them with asthma; any potential contribution would depend on general functions (such as transcriptional regulation or metabolism) that have not been described in asthmatic pathways.

Additionally, when we show a scatter plot of the principal components (Figure 4D), we can visually see the separation between the "health control" and "severe asthma" classes. Control group samples are highlighted with filled circles, while severe asthma samples are represented by empty circles.

When we analyze the expression levels of the genes proposed by the logistic regression model, we can see a clustering of genes that contribute positively (Figure 5A). However, we cannot see the same for genes that contribute negatively (Figure 5B).



**Figure 5. Expression levels for the top 20 most important genes according to the weight distribution (α) derived from the logistic regression model. Values were normalized by row. Details are available at Supplementary Table S2 and S3.**

# 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 underexplored in the literature. By applying a regularized logistic regression model to public gene expression data, we successfully recapitulated known asthma-related genes, such as *…* as a candidate of particular interest. Its relevance is supported by its high discriminative weight and by emerging literature on its involvement in immune processes linked to asthma. Its elevated expression in lung-resident dendritic cells and B cells, as documented in recent studies, further supports its putative role in adult-onset asthma and underscores the need for functional validation.

**Supplementary material:** The source code used is available at <https://github.com/LBS-UFMG/asthma_microarray>.

**Data availability:** The expression dataset from severe asthmatics, mild asthmatics, and healthy (GSE27011) controls is available in the Gene Expression Omnibus (GEO) and can be accessed at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27011.

**Funding:** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

**Acknowledgments:** The authors thank the funding agencies: CAPES; FAPEMIG; and CNPq.

**Conflict of Interest:** none declared.

# 5**. References**

[Abdelwahab, M.M., Al-Karawi, K.A., and Semary, H.E. 2023. Deep Learning-Based Prediction of Alzheimer’s Disease Using Microarray Gene Expression Data. *Biomedicines 11*, 12, 3304.](https://www.zotero.org/google-docs/?Gc1mlP)

[Agache, I. and Akdis, C.A. 2019. Precision medicine and phenotypes, endotypes, genotypes, regiotypes, and theratypes of allergic diseases. *The Journal of Clinical Investigation 129*, 4, 1493–1503.](https://www.zotero.org/google-docs/?Gc1mlP)

[Bigler, J., Boedigheimer, M., Schofield, J.P.R., et al. 2017. 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, 1311–1320.](https://www.zotero.org/google-docs/?Gc1mlP)

[Chatkin, J., Correa, L., and Santos, U. 2022. External Environmental Pollution as a Risk Factor for Asthma. *Clinical Reviews in Allergy & Immunology 62*, 1, 72–89.](https://www.zotero.org/google-docs/?Gc1mlP)

[Checkley, W. 2019. What have we learnt from studies on the association between urbanisation and asthma prevalence? *Thorax 74*, 11, 1011–1012.](https://www.zotero.org/google-docs/?Gc1mlP)

[Cremades-Jimeno, L., de Pedro, M.Á., López-Ramos, M., et al. 2021. Prioritizing Molecular Biomarkers in Asthma and Respiratory Allergy Using Systems Biology. *Frontiers in Immunology 12*.](https://www.zotero.org/google-docs/?Gc1mlP)

[Demšar, J., Curk, T., Erjavec, A., et al. 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research 14*, 71, 2349–2353.](https://www.zotero.org/google-docs/?Gc1mlP)

[Dweik, R.A., Boggs, P.B., Erzurum, S.C., et al. 2011. An Official ATS Clinical Practice Guideline: Interpretation of Exhaled Nitric Oxide Levels (FeNO) for Clinical Applications. *American Journal of Respiratory and Critical Care Medicine 184*, 5, 602–615.](https://www.zotero.org/google-docs/?Gc1mlP)

[Genç, M. 2024. Penalized logistic regression with prior information for microarray gene expression classification. *The International Journal of Biostatistics 20*, 1, 107–122.](https://www.zotero.org/google-docs/?Gc1mlP)

[Ghosh, D. 2002. Singular value decomposition regression models for classification of tumors from microarray experiments. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 18–29.](https://www.zotero.org/google-docs/?Gc1mlP)

[Goldin, J., Hashmi, M.F., and Cataletto, M.E. 2025. Asthma. In: *StatPearls*. StatPearls Publishing, Treasure Island (FL).](https://www.zotero.org/google-docs/?Gc1mlP)

[Hamid, Q., Tulic’, M.K., Liu, M.C., and Moqbel, R. 2003. Inflammatory cells in asthma: mechanisms and implications for therapy. *The Journal of Allergy and Clinical Immunology 111*, 1 Suppl, S5–S12; discussion S12-17.](https://www.zotero.org/google-docs/?Gc1mlP)

[Liu, X.-Y., Wu, S.-B., Zeng, W.-Q., Yuan, Z.-J., and Xu, H.-B. 2020. LogSum + L2 penalized logistic regression model for biomarker selection and cancer classification. *Scientific Reports 10*, 1, 22125.](https://www.zotero.org/google-docs/?Gc1mlP)

[Modena, B.D., Bleecker, E.R., Busse, W.W., et al. 2017. Gene Expression Correlated with Severe Asthma Characteristics Reveals Heterogeneous Mechanisms of Severe Disease. *American Journal of Respiratory and Critical Care Medicine 195*, 11, 1449–1463.](https://www.zotero.org/google-docs/?Gc1mlP)

[Mohammed, I., Elbashir, M.K., and Faggad, A.S. 2024. Singular Value Decomposition-Based Penalized Multinomial Regression for Classifying Imbalanced Medulloblastoma Subgroups Using Methylation Data. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 31*, 5, 458–471.](https://www.zotero.org/google-docs/?Gc1mlP)

[Mori, Y., Yokota, H., Hoshino, I., et al. 2021. Deep learning-based gene selection in comprehensive gene analysis in pancreatic cancer. *Scientific Reports 11*, 1, 16521.](https://www.zotero.org/google-docs/?Gc1mlP)

[Ni, R., Su, H., Burnett, R.T., Guo, Y., and Cheng, Y. 2024. 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.](https://www.zotero.org/google-docs/?Gc1mlP)

[Orsmark-Pietras, C., James, A., Konradsen, J.R., et al. 2013a. Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. *The European Respiratory Journal 42*, 1, 65–78.](https://www.zotero.org/google-docs/?Gc1mlP)

[Orsmark-Pietras, C., James, A., Konradsen, J.R., et al. 2013b. Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. *The European Respiratory Journal 42*, 1, 65–78.](https://www.zotero.org/google-docs/?Gc1mlP)

[Pearce, N., Douwes, J., and Beasley, R. 2000. Is allergen exposure the major primary cause of asthma? *Thorax 55*, 5, 424–431.](https://www.zotero.org/google-docs/?Gc1mlP)

[Price, D.B., Rigazio, A., Campbell, J.D., et al. 2015. Blood eosinophil count and prospective annual asthma disease burden: a UK cohort study. *The Lancet Respiratory Medicine 3*, 11, 849–858.](https://www.zotero.org/google-docs/?Gc1mlP)

[Skloot, G.S. 2016. Asthma phenotypes and endotypes: a personalized approach to treatment. *Current Opinion in Pulmonary Medicine 22*, 1, 3–9.](https://www.zotero.org/google-docs/?Gc1mlP)

[Vos, T., Lim, S.S., Abbafati, C., et al. 2020. 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. *The Lancet 396*, 10258, 1204–1222.](https://www.zotero.org/google-docs/?Gc1mlP)

[Wenzel, S.E. 2012. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nature Medicine 18*, 5, 716–725.](https://www.zotero.org/google-docs/?Gc1mlP)

[Zheng, X.-Y., Xu, Y.-J., Guan, W.-J., and Lin, L.-F. 2018. Regional, age and respiratory-secretion-specific prevalence of respiratory viruses associated with asthma exacerbation: a literature review. *Archives of Virology 163*, 4, 845–853.](https://www.zotero.org/google-docs/?Gc1mlP)