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

**Milenna Machado Pirovani 1,\*, [...], Raquel Melo-Minardi 1,† and Marcos Augusto dos Santos 1,†**

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

2D…

3D…

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. Genes with the highest absolute regression coefficients included CHI3L1, IL1RL1, and POSTN—well-documented markers of asthma—as well as HLA-DQA2, a gene with limited prior characterization in this context. The observed prominence of HLA-DQA2 supports emerging evidence of its involvement in adult-onset asthma, potentially through its immunological role in pulmonary antigen presentation. 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]. 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]. 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].

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.

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]. 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 .

Transcriptomic analyses have played a crucial role in biomarker discovery. The study by Modena et al. (2017) 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) 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) 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 specific to 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.

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, exceptionally 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) 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) 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., 2013]. 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]. 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].

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) [Ni; He; Chalise, 2023][Abdelwahab; Al-Karawi; Semary, 2023], such that:

A = T·S·Vᵀ, (1)

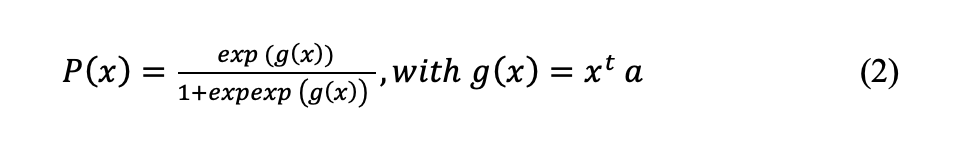
where **T** ∈ ℝᵐˣᵐ contains the left singular vectors (related to the samples); **S** ∈ ℝᵐˣⁿ is a rectangular matrix containing the singular values along the main diagonal and zeros elsewhere; **V** ∈ ℝⁿˣⁿ contains the right singular vectors (related to the attributes).

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

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]. The objective was to estimate a weight vector α ∈ ℝⁿ that defines the relative contribution of each attribute (gene) to the sample classification.

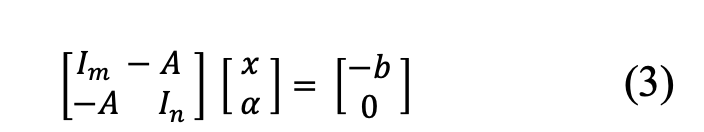
In this step, the *i*-th row of the transposed matrix A is represented by a vector x ∈ ℝP, where p is the number of attributes (genes) considered. The probability P(x) of a sample belonging to the positive class (*i.e.*, associated with the clinical condition of interest) is modeled by the logistic function:



(2)

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:



(3)

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]. The coefficients were sorted in ascending order, and the seven most negative and seven most positive values were selected, totaling 14 genes. Subsequently, the gene expression matrix A was reduced to a new matrix A∈ ℝmx14 , 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]. 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.

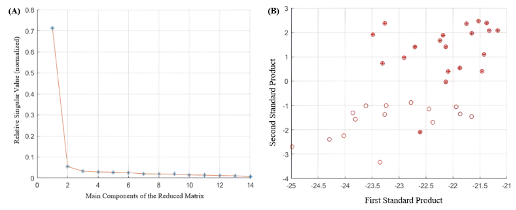
**Table 1. Results of logistic regression models built for datasets DS1 and DS2**

| **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 with 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).

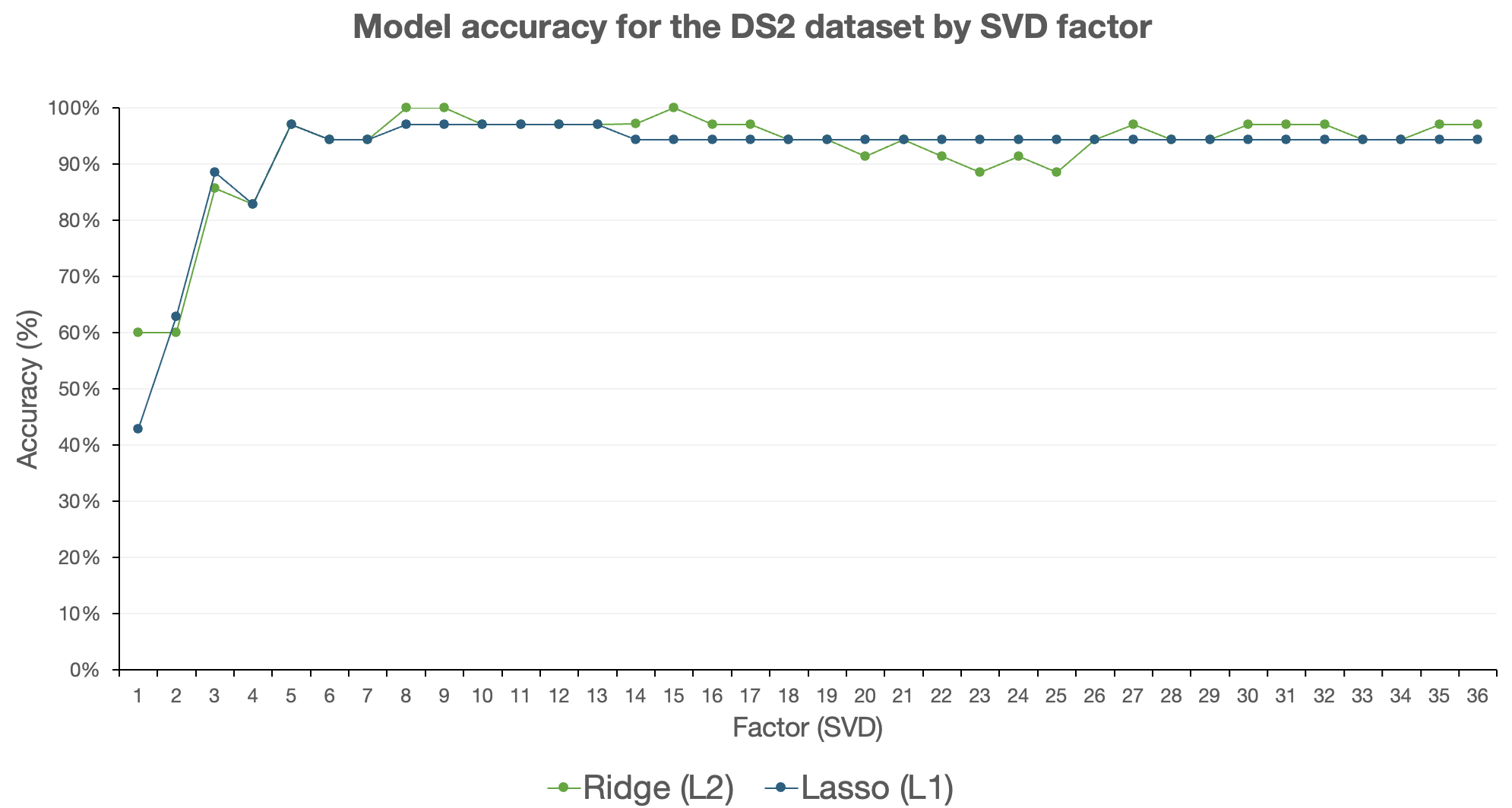
## 3.2 Noise reduction using SVD

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=36 and m=36). The factor plot indicated that only three dimensions of the auxiliary matrix could explain approximately 80% of the variance (Figure 1A). Additionally, when we show a scatter plot of the principal components (Figure 1B), we can visually see the separation between the "health control" and "severe asthma" classes.



**Figure 1. Importance of singular values by total factors (principal components) and two-dimensional scatter plot representing the inputs.** (A) Distribution of Relative Singular Values in the Reduced Matrix Normalized singular values of the reconstructed gene expression matrix are shown, using the 10 most relevant attributes selected by Modified Logistic Regression. (B) 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.

However, 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 three dimensions, the models achieve an accuracy close to 90%. After including five dimensions in the model, the accuracy remains high, ranging from 90% to 100%. Maximum accuracy can also be observed with the top nine and top 15 components.

## 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.

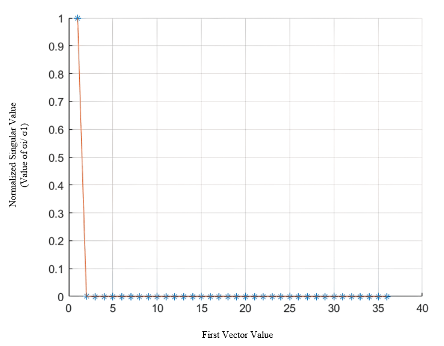
The logistic regression models outperformed the other algorithms for classification in all datasets evaluated (Table 2).

**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 | 0.943 | 0.667 | 0.882 | 0.941 |
| kNN | 0.741 | 0.829 | 0.743 | 0.462 | 0.786 | 0.690 |

## 3.2 Visualization of the Groups in the Reduced Space

Projection onto the first reduced space (Figure 1B) revealed partial spatial separation between controls and severe asthma cases, with overlapping clusters indicating limited class separability in the raw feature space. The Euclidean distance between group centroids was 2.96; however, substantial within-group dispersion underscores the need for dimensionality reduction and feature selection to enhance classification performance. This overlap suggests that many attributes do not significantly contribute to group differentiation, reinforcing the need for attribute selection to enhance model discriminative capacity.



**Figure 3. Distribution of the normalized singular values of matrix A ∈ ℝ³⁵ˣ²⁸²³¹. Each value represents the relative contribution of a principal component to the total variance of the data.**

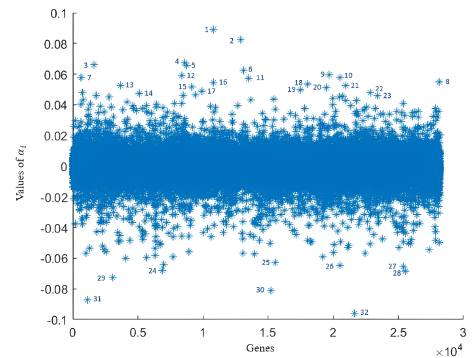


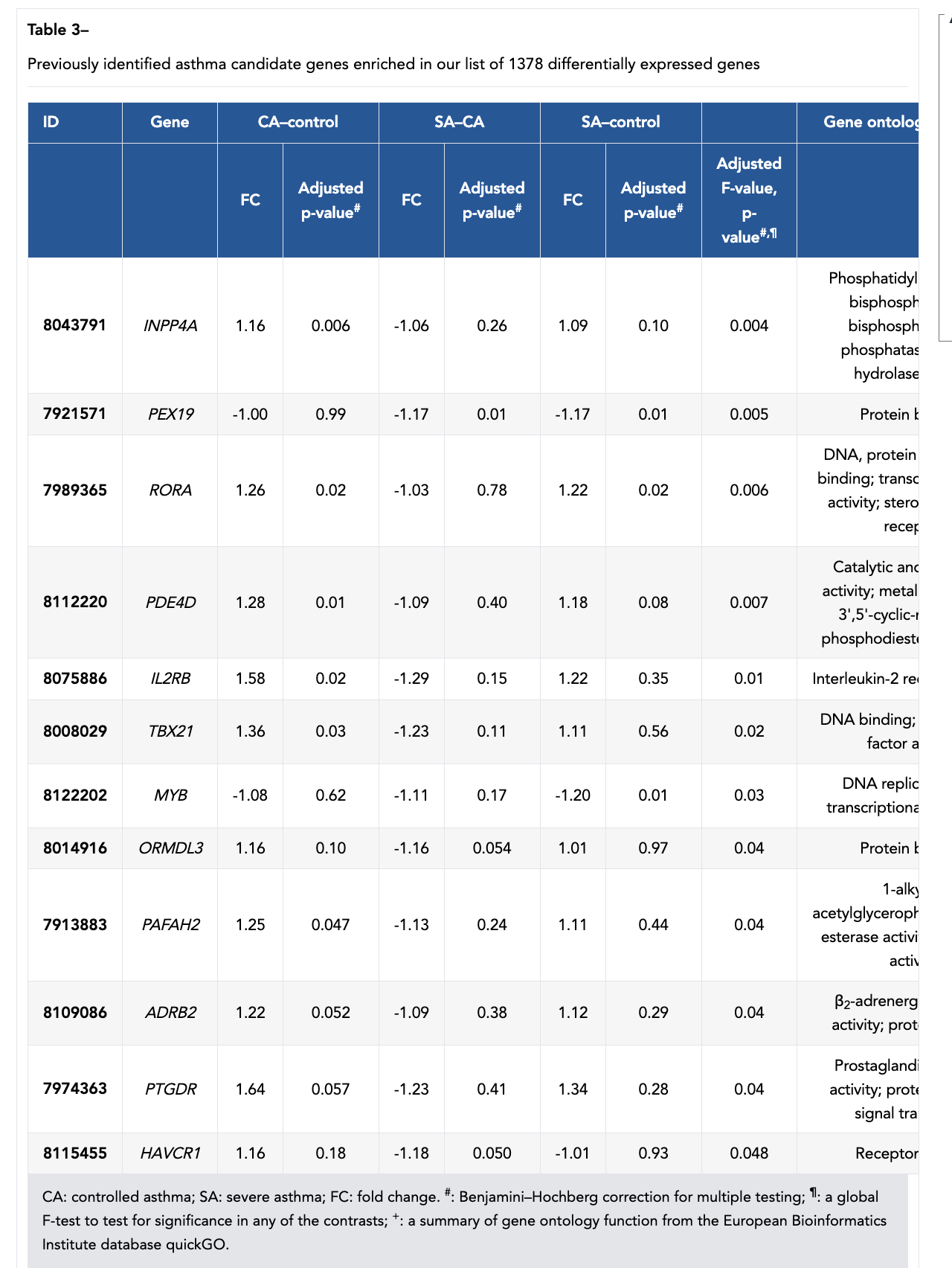
Figure 2. Projection of the Samples in the Space.

## 3.3 Distribution of Gene Weights after 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 3). A small subset exhibited high-magnitude weights, either positive or negative, suggesting their potential relevance for group discrimination. The ten genes with the highest absolute α values were selected for downstream analysis. Notably, multiple probes—such as 7981737 and 7914180—mapped to *IL1RL1,* reflecting probe redundancy in microarray design (Table 1). Functional annotations of the top candidates are provided in Table 2.

In (see Table 1)

| **Gene** | **Sonda** | **alpha** |
| --- | --- | --- |
| ORMDL3 | 8014916 |  |
| GSDMB | 8014903 |  |
| IL13 | 8107970 |  |
| IL4 | 7994280 |  |
| IL4R | 8107977 |  |
| STAT6 | 7964360 |  |
| IL1RL1 | 8044021 |  |
| IL33 | 8154295 |  |
| TSLP | 8107270 |  |
| IL18R1 | 8044035 |  |
| IL2RB | 8075886 |  |
| SMAD3 | 67487533 |  |
| IL6 | 8131803 |  |
| ADAM33 | 3663337 |  |
| HLA-DQ A1 | 8118548 |  |
| HLA-DQ B1 |  |  |
| HLA-DRB1 | 8125447 |  |



**Table 1. Key transcription clusters related to asthma identified in the current dataset**

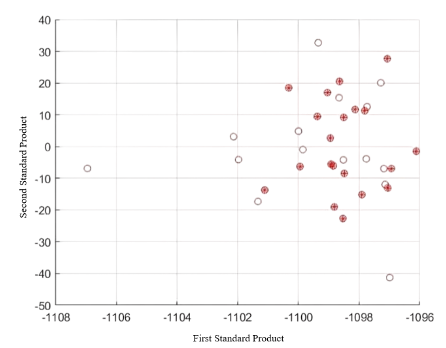
| **Positives** | | | **Negatives** | | |
| --- | --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **A** | **B** | **C** |
| 1 | 8007848 |  | 24 | 7967028 |  |
| 2 | 8029693 |  | 25 | 8055639 |  |
| 3 | 7914180 |  | 26 | 8105908 |  |
| 4 | 7983890 |  | 27 | 8154211 |  |
| 5 | 7985431 |  | 28 | 8155528 |  |
| 6 | 8031570 |  | 29 | 7928489 |  |
| 7 | 7903765 |  | 30 | 8052581 |  |
| 8 | 8180318 |  | 31 | 7908861 |  |
| 9 | 8098439 |  | 32 | 8117018 |  |
| 10 | 8106025 |  |  |  |  |
| 11 | 8035779 |  |  |  |  |
| 12 | 7981737 |  |  |  |  |
| 13 | 7934898 |  |  |  |  |
| 14 | 7948910 |  |  |  |  |
| 15 | 7990377 |  |  |  |  |
| 16 | 8007794 |  |  |  |  |
| 17 | 7998405 |  |  |  |  |
| 18 | 8081298 |  |  |  |  |
| 19 | 8075481 |  |  |  |  |
| 20 | 8096301 |  |  |  |  |
| 21 | 8110971 |  |  |  |  |
| 22 | 8129649 |  |  |  |  |
| 23 | 8135204 |  |  |  |  |

(A) Number in the Fig. 3; (B) Transcription cluster; (C) Gene Association. Probes marked with asterisks were not previously associated with asthma.

**Table 2. Key molecular associations and their functional roles were identified in the study. Probes marked with asterisks were not previously associated with asthma**

| **Association** | **Function** | **Citation** |
| --- | --- | --- |
| ***ALOX15*** | Enzyme involved in the biosynthesis of pro-inflammatory lipid mediators | [Boyce, 2022] |
| ***CCL26*** | Chemokine that attracts eosinophils to sites of inflammation | [Hoeck; Woisetschläger, 2001] |
| ***CHI3L1*** | Protein associated with inflammation and tissue remodeling | [Zhu et al., 2017] |
| ***CLCA1*** | Involved in mucus secretion and the inflammatory response in the airways | [Xu et al., 2022] |
| ***HLA-DBQ1*** | Part of the MHC class II complex, involved in the presentation of antigens to T helper cells | [Farina et al., 2019] [Gao et al., 2003] [Suarez-Pajes et al., 2021] |
| ***HLA-DQA1*** | Part of the MHC class II complex, involved in the presentation of antigens to T helper cells | [Farina et al., 2019] [Gao et al., 2003] [Suarez-Pajes et al., 2021] |
| ***HLA-DQA2\**** | Part of the MHC class II complex, involved in the presentation of antigens to T helper cells | [Clay et al., 2022] |
| ***IL13RA1*** | Receptor for interleukin-13, involved in the signaling of Th2-type immune responses | [Konstantinidis et al., 2007] |
| ***IL1RL1* (*ST2*)** | Receptor for interleukin-33 (IL-33), involved in the activation of Th2-type immune cells | [Gordon et al., 2016] |
| ***IL33*** | Cytokine that acts as an alarm of tissue damage, activating Th2-type immune cells | [Crim et al., 2022] |
| ***IL5RA*** | Alpha receptor for interleukin-5, crucial in the activation of eosinophils | [Cheong et al., 2005] [Elena-Pérez et al., 2021] |
| ***POSTN*** | ECM glycoprotein involved in tissue remodeling | [Burgess et al., 2021] |
| ***SERPINB2*** | Protease inhibitor, regulates inflammation and immune response | [ELBadawy; Abdel-Latif; El-Hady, 2017] |
| ***SPP1*** | Glycoprotein involved in cell adhesion and immune response | [Arjomandi et al., 2011] [Trinh et al., 2020] |
| ***TSLP*** | Cytokine involved in dendritic cell activation and allergic response | [Ebina-Shibuya; Leonard, 2023] |

Among the selected genes, *HLA-DQA2* stood out due to its high regression weight and limited prior characterization in asthma-related transcriptomic studies. Recent evidence suggests that *HLA-DQA2* expression is elevated in lung-resident B cells and dendritic cells (Schoettler et al. 2023), and may be associated with increased susceptibility to adult-onset asthma and asthma-related hospitalizations linked to MHC class II loci (Clay et al. 2022)(Yan et al. 2021). These findings support a potential immunogenic role for *HLA-DQA2* in asthma pathophysiology and justify further functional validation.



**Figure 3. Distribution of Weights Associated with the Attributes (Genes). These values indicate the relevance of each gene for distinguishing between the control and severe asthma groups.**

## 3.4 Classification by Logistic Regression Using All Attributes

The logistic regression model trained on the full expression matrix yielded a clear separation in log-odds values between control and severe asthma samples (Figure 4). However, the use of all 28,231 features raises the risk of overfitting, particularly given the limited sample size. This underscores the need for dimensionality reduction and attribute selection to improve model generalizability and interpretability.

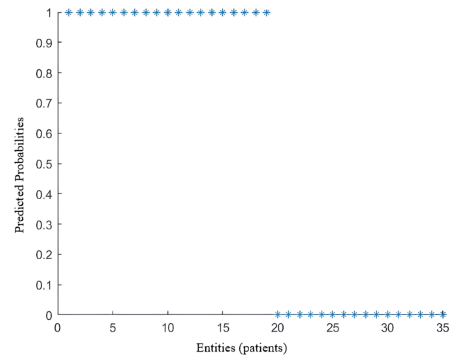


Figure 4. Classification Probabilities Obtained by Logistic Regression Using All Genes. Scores close to 1: control group; scores close to 0: severe asthma group.

## 3.5 Analysis of the Singular Values of the Reduced Matrix

Following the selection of the top ten genes, singular value decomposition was reapplied to the reduced matrix. The first singular value accounted for the majority of variance (Figure 5A), indicating that the selected features retained the core structure of the original dataset. This pattern is consistent with the presence of highly correlated transcriptional profiles among relevant genes and supports the utility of the selection strategy for dimensionality reduction without substantial information loss. The first singular value captures a substantial portion of the data variance, while subsequent values are markedly smaller, indicating that most information is concentrated in a few principal components. This pattern reflects highly correlated data and confirms that the attribute selection process effectively preserved the essential variability of the original dataset.

Control group samples are highlighted with filled circles, while severe asthma samples are represented by empty circles.

# 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 *IL1RL1* and *POSTN*, and identified *HLA-DQA2* 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.

<https://funcoup.org/uniprot/P01906/>

**Supplementary material:** The source code used is temporarily available at drive.google.com/file/d/1howYYl8ymIafTdfb7NcQISpyY7-rWP3e/, due to the peer review process.

**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: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES, Brazil; Fundação de Amparo à Pesquisa do Estado de Minas Gerais—FAPEMIG; and Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico—CNPq.

**Conflict of Interest:** none declared.

**7. References**

1. Hamid, Q.: Inflammatory cells in asthma: Mechanisms and implications for therapy. Journal of Allergy and Clinical Immunology 111(1), 5–17 (2003)

Hamid2003

2. Vos, Theo, Stephen S Lim, Cristiana Abbafati, 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): 1204–22.<https://doi.org/10.1016/S0140-6736(20)30925-9>.

3. Checkley, W.: What have we learnt from studies on the association between urbanisation and asthma prevalence? Thorax 74(11), 1011–1012 (2019)

4. Pearce, N., Douwes, J., Beasley, R.: Is allergen exposure the major primary cause of asthma? Thorax 55(5), 424–431 (2000)

5. Chatkin, J., Correa, L., Santos, U.: External Environmental Pollution as a Risk Factor for Asthma. Clinical Reviews in Allergy & Immunology 62, 72–89 (2022)

6. Ni, R., et al.: 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 (2024), https://doi.org/10.1016/j.oneear.2024.09.022

7. Zheng, X., et al.: Regional, age and respiratory-secretion-specific prevalence of respiratory viruses associated with asthma exacerbation: a literature review. Archives of Virology 163(4), 845–853 (2018), https://doi.org/10.1007/s00705-017-3700-y

8. Hashmi, M., Cataletto, M: Asthma. In: StatPearls Publishing (eds.) Treasure Island (FL) (2023)

9. Skloot, G. S.: Asthma phenotypes and endotypes. Current Opinion in Pulmonary Medicine 22(1), 3–9 (2016)

10. Porsbjerg, C., Menzies-Gow, A.: Co-morbidities in severe asthma: Clinical impact and management. Respirology 22(4), 651–661 (2017)

11. Kuruvilla, M. E., Lee, F. E., Lee, G. B.: Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease. Clinical Reviews in Allergy & Immunology 56, 219–233 (2019)

12. Cremades-Jimeno, L., et al.: Prioritizing Molecular Biomarkers in Asthma and Respiratory Allergy Using Systems Biology. Frontiers in Immunology 12(April), 1–22 (2021)

13. Ni, Y., He, J., Prabhakar Chalise: Randomized singular value decomposition for integrative subtype analysis of “omics data” using non-negative matrix factorization. Statistical Applications in Genetics and Molecular Biology 22(1) (2023)

14. Abdelwahab, M. M., Al-Karawi, K. A., Semary, H. E.: Deep Learning-Based Prediction of Alzheimer’s Disease Using Microarray Gene Expression Data. Biomedicines 11, 3304 (2023)

15. Genç, M.: Penalized logistic regression with prior information for microarray gene expression classification. The International Journal of Biostatistics 20(1), 107–122 (2022)

16. Liu, X. Y., et al.: LogSum + L2 penalized logistic regression model for biomarker selection and cancer classification. Scientific Reports 10, 22125, 1–16 (2020), https://doi.org/10.1038/s41598-020-79028-0

17. Mori, Y. et al.: Deep learning‑based gene selection in comprehensive gene analysis in pancreatic cancer. Scientific Reports 11(16521), 1–9 (2021)

18. Modena, B.D. et al.: Gene expression correlated with severe asthma characteristics reveals heterogeneous mechanisms of severe disease. American Journal of Respiratory and Critical Medicine 195(11), 1449–1463 (2017)

19. Farina, F. et al.: and are coordinately regulated. Cells 8(751), 1–14 (2019)

20. Gao, J. et al.: Association between HLA-DQA1, -DQB1 gene polymorphisms and susceptibility to asthma in northern Chinese subjects. Chinese Medical Journal 116(7), 1078–1082 (2003)

21. Suarez-Pajes, E.S. et al.: Targeted analysis of genomic regions enriched in African ancestry reveals novel classical HLA alleles associated with asthma in Southwestern Europeans. Scientific Reports 11(23686), 1–10 (2021)

22. Gordon, E.D. et al.: IL1RL1 asthma risk variants regulate airway type 2 inflammation. 1(14), 1–13 (2016)

23. Burgess, J.K. et al.: Periostin: contributor to abnormal airway epithelial function in asthma? European Respiratory Journal 57(2021), 1–12 (2021)

24. Xu, Y. et al.: CLCA1 mediates the regulatory effect of IL-13 on pediatric. Frontiers in Pediatrics, 1–9 (2022)

25. Elbadawy, N.E.; Abdel-Latif, R.S.; El-Hady, H.A.: Association between SERPINB2 gene expression by real-time PCR in respiratory epithelial cells and atopic bronchial asthma severity. Egyptian Journal of Immunology 24(1), (2017)

26. Hoeck, J.; Woisetschäger, M.: Activation of Eotaxin-3/CCL26 gene expression in human dermal fibroblasts is mediated by STAT6. The Journal of Immunology 167(6), 3216–3222 (2001)

27. Konstantinidis, A.K. et al.: Genetic association studies of interleukin-13 receptor α1 subunit gene polymorphisms in asthma and atopy. European Respiratory Journal 30(1), 40–47 (2007)

28. Boyce, J.A.: The role of 15-lipoxygenase 1 in asthma comes into focus. The Journal of Clinical Investigation 132(1), 1–3 (2022)

29. Zhu, Y. et al.: Association between risk of asthma and gene polymorphisms in CHI3L1 and CHIA: a systematic meta-analysis. BMC Pulmonary Medicine 17(193), 1–10 (2017)

30. Arjomandi, M. et al.: Polymorphism in osteopontin gene (SPP1) is associated with asthma and related phenotypes in a Puerto Rican population. Pediatric Allergy, Immunology, and Pulmonology 24(4) (2011)

31. Kim, H. et al.: Osteopontin contributes to late-onset asthma phenotypes in adult asthma patients. Experimental & Molecular Medicine 52, 253–265 (2020)

32. Cheong, H.S. et al.: Association analysis of interleukin 5 receptor alpha subunit (IL5RA) polymorphisms and asthma. Journal of Human Genetics 50(9), 628–634 (2005)

33. Elena-Pérez, S.; Heredero-Jung, D.H.; García-Sánchez, A.: Molecular analysis of IL-5 receptor subunit alpha as a possible pharmacogenetic biomarker in asthma. Frontiers in Medicine 7(February), 1–13 (2021)

34. Ebina-Shibuya, R.; Leonard, W.J.: Role of thymic stromal lymphopoietin in allergy and beyond. Nature Reviews Immunology 23(January) (2023)

35. Crim, C. et al.: IL-33 receptor inhibition in subjects with uncontrolled asthma: a randomized, placebo-controlled trial. Journal of Allergy and Clinical Immunology Global 1(4), 198–208 (2022)

36. Clay, S.M. et al.: Fine-mapping studies distinguish genetic risks for childhood- and adult-onset asthma in the HLA region. Genome Medicine 14(55), 1–16 (2022)

37. Schoettler, N., D. Decker, e A. I. Sperling. 2023. “Human Lung B Cells Express the Adult-onset Asthma Associated Non-classical HLA-DQ2 Genes and Proteins [abstract]”. *American Journal of Respiratory and Critical Care Medicine* 207: A1288.<https://doi.org/10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A1288>.

38. Yan, Q. et al.: A genome-wide association study of asthma hospitalizations in adults. Journal of Allergy and Clinical Immunology 147(3), 933–940 (2021)

**Referências corrigidas**

ABDELWAHAB, Mahmoud M.; AL-KARAWI, Khamis A.; SEMARY, Hatem E. Deep Learning-Based Prediction of Alzheimer’s Disease Using Microarray Gene Expression Data. **Biomedicines**, v. 11, n. 12, dec. 2023.

ARJOMANDI, Mehrdad *et al.* Polymorphism in Osteopontin Gene (SPP1) Is Associated with Asthma and Related Phenotypes in a Puerto Rican Population. **Pediatric allergy, immunology, and pulmonology**, v. 24, n. 4, p. 207–214, dec. 2011.

BOYCE, Joshua A. The role of 15 lipoxygenase 1 in asthma comes into focus. **The Journal of clinical investigation**, v. 132, n. 1, jan. 2022.

BURGESS, Janette K. *et al.* Periostin: contributor to abnormal airway epithelial function in asthma? **European Respiratory Journal**, v. 57, n. 2, 2021.

CHATKIN, Jose; CORREA, Liana; SANTOS, Ubiratan. External Environmental Pollution as a Risk Factor for Asthma. **Clinical reviews in allergy & immunology**, v. 62, n. 1, p. 72–89, feb. 2022.

CHECKLEY, William. What have we learnt from studies on the association between urbanisation and asthma prevalence? *In*: England: nov. 2019.

CHEONG, Hyun Sub *et al.* Association analysis of interleukin 5 receptor alpha subunit (IL5RA) polymorphisms and asthma. **Journal of Human Genetics**, v. 50, n. 12, p. 628–634, dec. 2005.

CLAY, Selene M. *et al.* Fine-mapping studies distinguish genetic risks for childhood- and adult-onset asthma in the HLA region. **Genome Medicine**, v. 14, n. 1, p. 55, may 2022.

CREMADES-JIMENO, Lucía *et al.* Prioritizing Molecular Biomarkers in Asthma and Respiratory Allergy Using Systems Biology. **Frontiers in immunology**, v. 12, p. 640791, 2021.

CRIM, Courtney *et al.* IL-33 receptor inhibition in subjects with uncontrolled asthma: A randomized, placebo-controlled trial. **The journal of allergy and clinical immunology. Global**, v. 1, n. 4, p. 198–208, nov. 2022.

DEMŠAR, Janez *et al.* Orange: Data Mining Toolbox in Python. **Journal of Machine Learning Research**, v. 14, p. 2349–2353, 2013.

EBINA-SHIBUYA, Risa; LEONARD, Warren J. Role of thymic stromal lymphopoietin in allergy and beyond. **Nature reviews. Immunology**, v. 23, n. 1, p. 24–37, jan. 2023.

ELBADAWY, Nissreen E.; ABDEL-LATIF, Randa S.; EL-HADY, Hoda A. Association between SERPINB2 Gene Expression by Real Time PCR in Respiratory Epithelial Cells and Atopic Bronchial Asthma Severity. **The Egyptian journal of immunology**, v. 24, n. 1, p. 165–181, jan. 2017.

ELENA-PÉREZ, Sandra *et al.* Molecular Analysis of IL-5 Receptor Subunit Alpha as a Possible Pharmacogenetic Biomarker in Asthma. **Frontiers in Medicine**, v. Volume 7-2020, 2021.

FARINA, Federica *et al.* HLA-DQA1 and HLA-DQB1 Alleles, Conferring Susceptibility to Celiac Disease and Type 1 Diabetes, are More Expressed Than Non-Predisposing Alleles and are Coordinately Regulated. **Cells**, v. 8, n. 7, jul. 2019.

GAO, Jinming *et al.* Association between HLA-DQA1, -DQB1 gene polymorphisms and susceptibility to asthma in northern Chinese subjects. **Chinese medical journal**, v. 116, n. 7, p. 1078–82, jul. 2003.

GENÇ, Murat. Penalized logistic regression with prior information for microarray gene expression classification. **The international journal of biostatistics**, v. 20, n. 1, p. 107–122, may 2024.

GOLDIN, Jennifer; HASHMI, Muhammad F.; CATALETTO, Mary E. **Asthma.** **StatPearls**Treasure Island (FL), jan. 2025.

GORDON, Erin D. *et al.* IL1RL1 asthma risk variants regulate airway type 2 inflammation. **JCI insight**, v. 1, n. 14, p. e87871, sep. 2016.

HAMID, Qutayba *et al.* Inflammatory cells in asthma: mechanisms and implications for therapy. **The Journal of allergy and clinical immunology**, v. 111, n. 1 Suppl, p. S5–S12; discussion S12-7, jan. 2003.

HOECK, J.; WOISETSCHLÄGER, M. Activation of eotaxin-3/CCLl26 gene expression in human dermal fibroblasts is mediated by STAT6. **Journal of immunology (Baltimore, Md. : 1950)**, v. 167, n. 6, p. 3216–22, sep. 2001.

KONSTANTINIDIS, A. K. *et al.* Genetic association studies of interleukin-13 receptor α1 subunit gene polymorphisms in asthma and atopy. **European Respiratory Journal**, v. 30, n. 1, p. 40–47, 2007.

KURUVILLA, Merin E.; LEE, F. Eun-Hyung; LEE, Gerald B. Understanding Asthma Phenotypes, Endotypes, and Mechanisms of Disease. **Clinical reviews in allergy & immunology**, v. 56, n. 2, p. 219–233, apr. 2019.

LIU, Xiao-Ying *et al.* LogSum + L2 penalized logistic regression model for biomarker selection and cancer classification. **Scientific Reports**, v. 10, n. 1, p. 22125, dec. 2020.

MODENA, Brian D. *et al.* Gene Expression Correlated with Severe Asthma Characteristics Reveals Heterogeneous Mechanisms of Severe Disease. **American journal of respiratory and critical care medicine**, v. 195, n. 11, p. 1449–1463, jun. 2017.

MORI, Yasukuni *et al.* Deep learning-based gene selection in comprehensive gene analysis in pancreatic cancer. **Scientific Reports**, v. 11, n. 1, p. 16521, aug. 2021.

NI, Ruijing *et al.* 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**, v. 7, n. 11, p. 1953–1969, 2024.

NI, Yonghui; HE, Jianghua; CHALISE, Prabhakar. Randomized singular value decomposition for integrative subtype analysis of ‘omics data’ using non-negative matrix factorization. **Statistical Applications in Genetics and Molecular Biology**, v. 22, n. 1, p. 1–14, jan. 2023.

ORSMARK-PIETRAS, Christina *et al.* Transcriptome analysis reveals upregulation of bitter taste receptors in severe asthmatics. **The European respiratory journal**, v. 42, n. 1, p. 65–78, jul. 2013.

PEARCE, N.; DOUWES, J.; BEASLEY, R. Is allergen exposure the major primary cause of asthma? **Thorax**, v. 55, n. 5, p. 424–31, may 2000.

PORSBJERG, Celeste; MENZIES-GOW, Andrew. Co-morbidities in severe asthma: Clinical impact and management. **Respirology (Carlton, Vic.)**, v. 22, n. 4, p. 651–661, may 2017.

SCHOETTLER, N.; DECKER, D.; SPERLING, A. I. Human Lung B Cells Express the Adult-onset Asthma Associated Non-classical HLA-DQ2 Genes and Proteins [abstract]. *In*: 2023. Available at: <<https://doi.org/10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A1288>>

SKLOOT, Gwen S. Asthma phenotypes and endotypes: a personalized approach to treatment. **Current opinion in pulmonary medicine**, v. 22, n. 1, p. 3–9, jan. 2016.

SUAREZ-PAJES, Eva *et al.* Targeted analysis of genomic regions enriched in African ancestry reveals novel classical HLA alleles associated with asthma in Southwestern Europeans. **Scientific Reports**, v. 11, n. 1, p. 23686, dec. 2021.

TRINH, Hoang Kim Tu *et al.* Osteopontin contributes to late-onset asthma phenotypes in adult asthma patients. **Experimental & Molecular Medicine**, v. 52, n. 2, p. 253–265, feb. 2020.

VOS, Theo *et al.* 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**, v. 396, n. 10258, p. 1204–1222, 17 oct. 2020.

XU, Yanan *et al.* CLCA1 mediates the regulatory effect of IL-13 on pediatric asthma. **Frontiers in Pediatrics**, v. Volume 10-2022, 2022.

YAN, Qi *et al.* A genome-wide association study of asthma hospitalizations in adults. **The Journal of allergy and clinical immunology**, v. 147, n. 3, p. 933–940, mar. 2021.

ZHENG, Xue-yan *et al.* Regional, age and respiratory-secretion-specific prevalence of respiratory viruses associated with asthma exacerbation: a literature review. **Archives of Virology**, v. 163, n. 4, p. 845–853, apr. 2018.

ZHU, Yanting *et al.* Association between risk of asthma and gene polymorphisms in CHI3L1 and CHIA: a systematic meta-analysis. **BMC Pulmonary Medicine**, v. 17, n. 1, p. 193, dec. 2017.