Title: Identification of Novel Genes Associated with Chronic Obstructive Pulmonary Disease Using Different Penalized Logistics Regression with High-Dimensional Biological Data from Human Sputum Cells

Kimiya Gohari[1], Anoshirvan Kazemnejad[1*], Shayan Mostafaei[2], Ali Sheidaei[3], Maryam S Daneshpour[4], Mahdi Akbarzadeh[4], Farshad Sharifi[5]

1. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran.

2. Department of Biostatistics, School of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran

3. Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.

4. Cellular and Molecular Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

5. Elderly Health Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran

* Corresponding author: Anoshirvan Kazemnejad: Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, P.O. BOX 14115-111, Tehran, Iran.

Abstract

Background*:*

Comparison of LASSO, smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) logistic classifiers in order to reconnaissance of related genes with COPD disease and assessing the genes effects on the progression of the disease based on one of the main classes of cells involved in the disease, Sputum Cells.

We used a genome-wide expression profiling to define gene networks relevant to the disease. The data retrieved from Gene Expression Omnibus (GEO) with accession numbers "GSE22148". From 143 samples in GOLD stage 2-4 COPD ex-smokers, 54,675 probes primary were assessed. After normalization, LASSO, SCAD and MCP logistic regressions were applied. K-fold cross-validation scheme was used to evaluate the performance of two methods. All of the computational processes were done using "ncvreg", "Affy," "Limma" and "SVA" R packages.

Results:

The results of LASSO (AUC=0.95, sensitivity= 0.91, specificity= 0.86) and SCAD (AUC=0.97, sensitivity= 0.95, specificity= 0.85) logistic regression were almost similar. There were 23 and 22 significantly associated genes for LASSO and SCAD, respectively. The only difference between these models is related to "*stromal interaction molecule 2*". Comparing to MCP approach, the most conservative method, we detected only 7 significant genes (AUC= 0.94, sensitivity= 0.94, specificity= 0.82).

Conclusions:

In the present study, the relative expressions of thousands of the genes were assessed and identified as associated genes with the progression of COPD. Differential analysis of gene expression data is able to reduce the number of genes but in a limited manner. In order to find an efficient and small subset of genes, we should use alternative approaches like logistic regression. Regularization solves the high dimensionality problem in using this kind of regression.

Keywords: COPD, Gene expression, Penalized Statistical Modeling, Lasso, SCAD, MCP, Regularization

1. **Background**

Chronic obstructive pulmonary disease (COPD) is a progressive inflammatory disease mostly characterized by airway obstruction and is predicted to be among the first three causes of death worldwide by 2030 (1). General exhibitions include emphysema, small airway obstructions, and chronic bronchitis. In a healthy airway and alveolar development background, smoking is the first-line risk factors followed closely by the history of maternal/paternal asthma, maternal smoking, and childhood asthma or respiratory infections. Polluted air, second-hand smoking, and malnutrition could also lead to COPD in the susceptible population (2).

COPD is a complex disease and besides the environmental factors, the disease course is dependent on interactions between different genes. DNA microarrays now permit scientists to screen thousands of genes simultaneously and determine whether the expression pattern of these genes changes in normal and COPD tissue. So, new analytical methods must be developed for selecting genes related to COPD. Since there are numerous allele variants involved in COPD, single nucleotide polymorphism (SNP) has been vastly used and shown numerous susceptibility regions on the genome (3, 4). Therefore, the associations between the genomics and the disease incidence and progression could be studied more precisely through machine-learning techniques (5). Also, network medicine has been introduced for facilitating the investigations on genomics, transcriptomics, proteomics, and other "–omics" to cast a more elucidating light on the complexity of the pathogenesis of diseases like COPD (6). One of the properties of microarray data is that the number of genes (p) exceeds the number of samples (n). They are dealing with the situation, which is commonly known as the high dimensional dataset. However, logistic

regression as a highly appropriate classification tool for such high dimensional datasets from the microarray technique has a few drawbacks, such as the emergence of irrelevant data (7, 8).

Moreover, regression analysis has been established to tend to overlook the multicollinearity problem (strong correlation between two or more than two genes in the regression model) (9). So, overfitting and multicollinearity are the most common problems that arise in high-dimensional data when applying statistical classification and prediction methods (10). Nowadays, researchers update, improve, and optimize the models such as the LASSO, minimax concave penalty (MCP), and smoothly clipped absolute deviation (SCAD) to introduce statistical learning models to overcome this issue. Penalized Logistic Regression models represent spares and interpretable model in high dimensional datasets and control the multicollinearity (7, 8). Up until now, there has been no study on -omics data integration on COPD with these approaches.

Although LASSO has many excellent properties, it is a biased estimator, and this bias does not necessarily go away as increases. The bias of the LASSO estimate for a truly non-zero variable is constant even for large regression coefficients. One approach to reducing the bias of the LASSO is to use the weighted penalty approach. If we choose the weights in such a way that the variables with large coefficients have smaller weights, then we can reduce the estimation bias of the LASSO. It is the motivation of adaptive LASSO approaches. The SCAD penalty retains the penalization rate (and bias) of the LASSO for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases. The idea behind the MCP is very similar. In comparison to SCAD, however, the MCP relaxes the penalization rate immediately while with the SCAD, the rate remains flat for a while before decreasing.

On the other hand, from a biological perspective, only a small subset of genes is strongly indicative of the target disease, and most genes are irrelevant to COPD classification and prediction (5). To fill this gap, the present study designed to apply statistical-learning methods for better understanding the genetic etiology in COPD affected by smoking habit. LASSO, MCP, and SCAD logistic regression applied to identify the most important genes related to GOLD stage 2-4 COPD ex-smokers.

2. *Results*

2.1. *Differential analysis of genes expression data*

Differential analysis was performed on the expression profiling of 54675 probes by the array. The expression profilings of the probes were extracted from 143 patients in GOLD stage 2-4 COPD. The results of the differential analysis showed significant expressions for the top 250 genes after adjusting p values by the Benjamini-Hochberg-FDR correction at α =0.05 (More details shown in additional file 1).

2.2. *Gene selection and prediction*

Based on the LASSO logistic regression, 24 genes significantly affect COPD severity (Table 1). The results of MCP and SCAD models are also presented in Table 1. The MCP model is the most conservative one and shows only seven genes as significant. The results of SCAD and LASSO models were similar except for one gene, "stromal interaction molecule 2", which was not significant in the SCAD model (p-value=0.074). While SCAD logistic classifier had the highest AUC

(0.97, 95% CI=0.95-0.98), sensitivity (95%), specificity (85%), and Youden index (0.81), LASSO logistic classifier had AUC (0.95, 95% CI=0.92-0.96), sensitivity (91%), specificity (86%) and Youden index (0.78). The ROC curves for the three approaches are depucted in figure 1. The difference between LASSO and SCAD accuracy indices was not significant (p-value=0.39). Conversely, MCP logistic classifier showed the lowest AUC (0.94, 95% CI=0.91-0.95), sensitivity (94%), specificity (82%) and Youden index (0.76). Based on the SCAD logistic regression results, the most important selected genes were RNF130, SLC38A2, STX6, PLCB1, LOC102724382, ZNF33A, LOC100288675, ESYT2, LARP4B, CACNA1G, LOC100507634, TAF15, LINC00693, TMEM182, PRDX2, PELP1, LAMA1, RPIA, and AMOTL1.

Consequently, 24 candidate genes identified here were associated with the progression of COPD by these classifiers. Based on their patterns of co-expression for twenty-four candidate genes by the heatmap plot for hierarchical clustering showed that all of the candidate genes were divided into the two major clusters (Figure 2).

Figure 3 summarizes the results' overlaps of the gene selection between these classifiers. According to this figure, the MCP model, in contrast to other models, showed the "*Caldesmon 1*" as the only significant gene in this classifier.

3. *Discussion*

Chronic Obstructive Pulmonary Disease (COPD) is a progressive life-threatening lung disease that causes breathlessness, and it was the fifth leading cause of death in 2002 (11). This disease continues to be a significant cause of morbidity, mortality, and health-care costs worldwide (12). Total deaths from COPD are projected to increase by more than 30% in the next ten years unless urgent action is taken to reduce the underlying risk factors, especially tobacco consumption. The global burden of the disease study reports a prevalence of 251 million cases of COPD in 2016, worldwide. Estimates show that COPD becomes the third leading cause of death in 2017, worldwide. Recent studies have indicated that the occurrence of lung cancer is a multiple-factors and multiple-step process, and it is the result of interaction between genetic and environmental exposure factors (13). COPD is a complex disease that is influenced by genetic factors, environmental influences, and genotype-environment interactions. Genotype-environment interactions have been of interest to geneticists for decades (14). The main risk factors for COPD included smoking, indoor air pollution, outdoor air pollution, genetic factors, and occupational dust and chemicals. Some cases of COPD are due to long-term asthma.

In this study, we also identified 24 significant genes associated with COPD progression. That may represent novel biomarkers in the prognosis of COPD. In our analyses, the most significantly regulated novel genes were: RNF130, SLC38A2, STX6, PLCB1, LOC102724382, ZNF33A, LOC100288675, ESYT2, LARP4B, CACNA1G, LOC100507634, TAF15, LINC00693, TMEM182, PRDX2, PELP1, LAMA1, RPIA, STIM2, AMOTL1, and CALD1 based on the Z scores.

Stromal interaction molecules, STIM2, is a regulator of store-operated calcium (Ca2+) entry as well as basal cytoplasmic Ca2+ levels in human cells (15). As is known, E2 exposure inhibited STIM1 translocation in airway epithelia, preventing SOCE. E2 can signal non genomically by inhibiting basal phosphorylation of STIM1, and STIM2, leading to a reduction in SOCE (16). Another study showed that STIM1 and STIM2 were significant as up-regulated genes versus healthy controls and healthy smokers (17). Also, AMOTL1 via the activation of LKB1/AMPK signaling and IFN-γ-induced hyperpermeability of cultured human lung microvascular endothelial cells by maintaining the levels of AmotL1 is related to lung function (18). Angiomotin Like one and caldesmon, one both have a role in involvement in Adhesion and Cell Motility in lung airway and alveolars and may have a role in obstruction of the airway by a problem in the expulsion of produced mucosa and destruction of alveolar walls or spasm in small airways (19). The STIM2 and AMOTL1 were selected as the most important genes in this study may reveal these genes as a novel target in the treatment of COPD (20). Also, caldesmon one gene (CALD1) as a novel gene associated with both the overall survival and the disease-free survival in bladder cancer patients (21), diabetic nephropathy (22, 23), and glioma neovascularization (24). RNF130, ring finger protein 130, is a candidate gene for Endoplasmic reticulum-associated degradation using phylogenetic tree analysis (25). RNF130 is involved in the pathogenesis of gestational trophoblastic diseases (26). CALD1 and RNF130 genes are not previously detected in COPD studies that may represent novel biomarkers in the diagnosis or prognosis of COPD. However, STX6 is involved in diverse cellular functions in a variety of cell types and has been shown to regulate many intracellular membrane trafficking events such as endocytosis, recycling, and anterograde and retrograde trafficking. The oncogenic roles of STX6 in the progression of

esophageal squamous cell carcinoma (ESCC) is established, and it might be a valuable target for ESCC therapy (27). An Epigenome-Wide Association Study found that STX6 may one of genes that associated with atopic asthma. They also reported that STX6 may has a role in methylate process that is seen in this disease (28)

In other studies, the association between PLCB1 and the decrease of FEV1 and FEV1/FVC in their participation was shown (29). Another study reported that PLCB1 at least in two pedigree with severe COPD was observed (30) Other group found an association between ZNF33A gen and remission from asthma(31) LARP4B is a target gene of miR106 inhibition of miR-106b that suppressed the mRNA and protein expression of cancer-related genes (32). Syntaxin 6 may have a role in regulating neutrophil secondary granule exocytosis also stimulation of cells by Ca2+. This role may be effective in the inflammatory process that results in obstruction in airways in COPD (33).

Chronic obstructive pulmonary disease is a progressive health problem that is accompanied by dyspnea, cough, and sputum production. Dyspnea is caused by two mechanisms: 1) alveolar cell destruction and inability to the alveolar wall to maintain their structure and decrease available respiratory gases exchange surface area, and 2) inflammation of airways that causes narrowing of small airways, and this can result in a problem in the passing of air in the small airways. Several molecular pathophysiologic pathways induce similar clinical symptoms and signs, such as limitations in pulmonary function and caught. The studies were shown that chronic inflammation and an increase of oxidative stress by smoking might have a role in the progression of COPD. The inflammatory cells could release mediators, such as proteases and cytokines; these mediators may contribute to tissue remodeling. Chemoattractant factors, chemokines, and attract

additional inflammatory cells to pulmonary tissue: epithelial and proinflammatory cytokines, chemokines, and other mediators (34).

It seems that two main mechanisms participate in the development of COPD, and several genes may involve in these processes. The first mechanism is oxidative stress and response of immune cells such as neutrophils, CD4, and CD8 lymphocytes and macrophages, which have an important role in this inflammatory process. It is reported that macrophage 5- 10 times increased in airways, lung parenchyma, BAL fluid, and sputum in patients with COPD (35). Gens such as AMOTL1, Syntaxin 6, and PRX2 may have a role in inflammation or oxidative stress response. Some other genes, such as Cacna1g and smooth muscles that existed in small airways. Some genes, such as CALD1, may have a role in the maintenance of cell members and may in the destruction of alveolars walls. And another mechanism may induce by these genes may production of glycoprotein s and amyloids that help to obstruct the small airways.

Regularization-based logistic regression models (e.g., LASSO, MCP, and SCAD) have already been used widely in microarray analysis (36). In this study, based on 10-fold cross-validation, the SCAD and LASSO regularized logistic regression models were found to perform better than the MCP logistic regression. LASSO has satisfying properties, and it is good for simultaneous estimation and eliminating trivial genes but is not good for grouped selection in microarray data. Our previous study showed that LASSO-based methods' (e.g., elastic-net) penalty are useful for gene selection in microarray COPD data (37). However, it is known that LASSO requires rather stringent conditions on the design matrix to be variable selection consistent (38). Non-convex penalized high-dimensional regression has recently received considerable attention to focus on identifying

the unknown sparsity pattern,. We recommended the SCAD penalty, which enjoys the oracle property for variable selection. In this study, SCAD regularized logistic regression models was found to perform better than LASSO. The performance of the MCP procedure is satisfactory, but its optimal performance depends on the tuning parameter (39). In this study, MCP is the most conservative method for gene selection. In microarray data classification, SCAD regularized logistic regression may provide a useful methodology for future studies in the discovery of novel diagnostic- and prognostic biomarkers and novel therapeutic targets in the treatment of COPD.

4. *Conclusion*

In the present study, the relative expressions of thousands of the genes were assessed and identified as associated genes with the progression of COPD. Differential analysis of gene expression data is able to reduce the number of genes but in a limited manner. In order to find an efficient and small subset of genes, we should use alternative approaches like logistic regression. Regularization solves the high dimensionality problem in using this kind of regression.

In this dataset, SCAD logistic regression, with a lot of advantages in theoretical, computational, and practical aspects, had a higher accuracy rate than LASSO and MCP penalties and might be useful for diagnosis or suitable intervention for COPD.

5. *Methods*

5.1. *Study population and dataset*

The raw data of gene expression architecture in the small airway epithelium (SAE) of COPD retrieved from the Gene Expression Omnibus (GEO) site in the National Center of Biotechnology Information (NCBI) database (3), with the accession number "GSE22148", with 54,675 probes from 143 patients in GOLD stage 2-4 COPD. Genome-wide gene expression analysis performed using Affymetrix Human Genome U133 Plus 2.0 array (GPL570) (40).

In that study, two sputum studies were performed in GOLD stage 2-4 COPD ex-smokers from the ECLIPSE cohort. First, gene array profiling at baseline for 1480 patients was performed. One year later, samples from a separate population of 176 patients were used for real-time PCR. The gene expression findings for IL-18R were further analyzed using immunohistochemistry in lung tissue and induced sputum samples from patients outside the ECLIPSE cohort.

5.2. *Normalization and filtering of primary probes*

The "sva" and "affy" packages were used respectively for removing batch effects and other unwanted variations in data and for statistical comparisons (41, 42). Also, the standardization and normalization in the "limma" package performed (43). In addition, differential analysis of genes expression data was conducted using the adjusted P-value based on the Benjamini-Hochberg-FDR correction at $\alpha = 0.05$. All statistical analyses performed using R version 3.5.2 (44).

5.3. *LASSO, MCP, and SCAD logistic regressions*

The ridge regression adds squared magnitude of coefficients as penalty term to regularize parameters. All of the estimated coefficients are non-zero, so no gene selection is performed. However, LASSO regression uses the absolute value of magnitude of coefficient as penalty term instead (Equation 1), and hence provide automatic gene selection. On the other hand, the ridge penalty tends to shrink the coefficients of correlated variables toward each other, which is good for multicollinearity and grouped selection. However, the LASSO penalty is somewhat indifferent to the choice among a set of strong but correlated variables. Therefore, LASSO is good for simultaneous estimation and eliminating trivial genes but not for grouped selection. However, it is known that LASSO requires rather stringent conditions on the design matrix to be variable selection consistent (38).

$$P_{Lasso}(x|\lambda) = \lambda|x| \qquad \text{(Equation 1)}$$

Non-convex penalized high-dimensional regression has recently received considerable attention, especially for identifying the unknown sparsity pattern. Fan and Li recommended the SCAD penalty, which enjoys the oracle property for variable selection (45). It can estimate the zero coefficients as exact zero with probability approaching one and estimate the non-zero coefficients as efficiently as if the true sparsity pattern is known in advance (Equation 2) (46).

$$P_{SCAD}(x|\lambda.\gamma) = \begin{cases} \lambda|x| & if\ |x| \le \lambda \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma - 1)} & if\ \lambda < |x| < \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & if\ |x| \ge \gamma\lambda \end{cases} \qquad \text{(Equation 2)}$$

Zang proposed MCP and devised a novel PLUS algorithm which, when used together, can achieve the oracle property under certain regularity conditions (Equation 3) (47). The mentioned logistic classifiers were done by "ncvreg" R packages (48).

$$P_{MCP}(x;\lambda) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma} & if\ |x| \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & if\ |x| > \gamma\lambda \end{cases} \qquad \text{(Equation 3)}$$

5.4. *Cross-validation, Stability, and Accuracy*

K-fold cross-validation scheme (K-CV) is a very commonly employed technique used to evaluate classifier performance. K-CV estimation of the error is the average value of the errors committed in each fold. Thus, the K-CV error estimator depends on two factors: the training set and the partition into folds (49). In the present study, the algorithms split the data set by using 100 times repeated random sub-sampling in 10-fold cross-validation, permuting the sample labels every time. The cross-validated performance is summarized by observed sensitivity and specificity and the Youden index. Furthermore, the area under the Receiver Operator Characteristic (ROC) curve (AUC) was used to calculate the accuracy of classifiers' performance (50, 51). We used “cv.ncvreg” and “roc” function in “ncvreg” and “pROC” R packages (52) for K-CV and ROC analysis respectively.

5.5. *Interactive Cluster Heatmap*

A heatmap is a popular graphical method for visualizing high-dimensional data. Rows and columns are sorted using a hierarchical clustering technique (53). The interactive cluster heatmap was applied by the "heatmaply" R package (54).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the Gene Expression Omnibus (GEO) site in the National Center of Biotechnology Information (NCBI) database, with the accession number "GSE22148" repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22148

Competing interests

The authors declare that they have no competing interests.

Funding

Funding was provided by National Institute for Medical Research Development.

Authors' contributions

KG and ASH performed the analysis. AK and SHM designed the analysis. AK, KG, ASH wrote the manuscript with support from MSD, MA, and FSH. All authors read and approved the final manuscript.

Acknowledgment

The authors would like to thank the National Institute for Medical Research Development for supporting this project; Grant No. 982759.

References

1. Zhao J, Li M, Chen J, Wu X, Ning Q, Zhao J, et al. Smoking status and gene susceptibility play important roles in the development of chronic obstructive pulmonary disease and lung function decline: A population-based prospective study. Medicine (Baltimore). 2017;96(25):e7283.
2. Postma DS, Bush A, van den Berge M. Risk factors and early origins of chronic obstructive pulmonary disease. Lancet. 2015;385(9971):899-909.
3. Liu Y, Huang K, Wang Y, Hu E, Wei B, Song Z, et al. Integration of SNP Disease Association, eQTL, and Enrichment Analyses to Identify Risk SNPs and Susceptibility Genes in Chronic Obstructive Pulmonary Disease. BioMed Research International. 2020;2020.
4. Ortega-Martínez A, Pérez-Rubio G, Ambrocio-Ortiz E, Nava-Quiroz KJ, de Jesus Hernández-Zenteno R, Abarca-Rojano E, et al. The SNP rs13147758 in the HHIP Gene Is Associated With COPD Susceptibility, Serum, and Sputum Protein Levels in Smokers. Frontiers in genetics. 2020;11.
5. Hardin M, Silverman EK. Chronic Obstructive Pulmonary Disease Genetics: A Review of the Past and a Look Into the Future. Chronic Obstr Pulm Dis. 2014;1(1):33-46.
6. Silverman EK, Loscalzo J. Network medicine approaches to the genetics of complex diseases. Discov Med. 2012;14(75):143-52.
7. Chai H, Liang Y, Liu XY. The L(1/2) regularization approach for survival analysis in the accelerated failure time model. Comput Biol Med. 2015;64:283-90.
8. Huang HH, Liu XY, Liang Y. Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L1/2 +2 Regularization. PLoS One. 2016;11(5):e0149675.
9. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology (Sunnyvale). 2016;6(2).
10. Cui Y, Zheng C-H, Yang J, Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Computers in biology and medicine. 2013;43(7):933-41.
11. Chan T-C, Wang H-W, Tseng T-J, Chiang P-H. Spatial clustering and local risk factors of chronic obstructive pulmonary disease (COPD). International journal of environmental research and public health. 2015;12(12):15716-27.
12. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. The Lancet. 2007;370(9589):765-73.
13. Wang Z, Feng F, Zhou X, Duan L, Wang J, Wu Y, et al. Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. Oncotarget. 2017;8(55):94793.
14. Sandford A, Silverman E. Chronic obstructive pulmonary disease• 1: Susceptibility factors for COPD the genotype–environment interaction. Thorax. 2002;57(8):736-41.
15. Stathopulos PB, Zheng L, Ikura M. Stromal interaction molecule (STIM) 1 and STIM2 calcium sensing regions exhibit distinct unfolding and oligomerization kinetics. Journal of Biological Chemistry. 2009;284(2):728-32.
16. Sheridan JT, Gilmore RC, Watson MJ, Archer CB, Tarran R. 17β-Estradiol Inhibits Phosphorylation of Stromal Interaction Molecule 1 (STIM1) Protein IMPLICATION FOR STORE-OPERATED CALCIUM ENTRY AND CHRONIC LUNG DISEASES. Journal of Biological Chemistry. 2013;288(47):33509-18.
17. Deng F, Dong H, Zou M, Zhao H, Cai C, Cai S. Polarization of neutrophils from patients with asthma, chronic obstructive pulmonary disease and asthma-chronic obstructive pulmonary disease overlap syndrome. Zhonghua yi xue za zhi. 2014;94(48):3796-800.
18. Suzuki R, Nakamura Y, Chiba S, Mizuno T, Abe K, Horii Y, et al. Mitigation of tight junction protein dysfunction in lung microvascular endothelial cells with pitavastatin. Pulmonary pharmacology & therapeutics. 2016;38:27-35.

19. Mercer BA, Lemaître V, Powell CA, D'Armiento J. The epithelial cell in lung health and emphysema pathogenesis. Current respiratory medicine reviews. 2006;2(2):101-42.
20. Nakajima Y, Nakamura Y, Shigeeda W, Tomoyasu M, Deguchi H, Tanita T, et al. The Role of Tumor Necrosis Factor-α and Interferon-γ in Regulating Angiomotin-Like Protein 1 Expression in Lung Microvascular Endothelial Cells. Allergology International. 2013;62(3):309-22.
21. Liu Y, Wu X, Wang G, Hu S, Zhang Y, Zhao S. CALD1, CNN1, and TAGLN identified as potential prognostic molecular markers of bladder cancer by bioinformatics analysis. Medicine. 2019;98(2).
22. Wang Z, Wang Z, Zhou Z, Ren Y. Crucial genes associated with diabetic nephropathy explored by microarray analysis. BMC nephrology. 2016;17(1):128.
23. Śnit M, Nabrdalik K, Długaszek M, Gumprecht J, Trautsolt W, Górczyńska-Kosiorz S, et al. Association of rs 3807337 polymorphism of CALD1 gene with diabetic nephropathy occurrence in type 1 diabetes—preliminary results of a family-based study. Endokrynologia Polska. 2017;68(1):13-7.
24. Zheng P-P, Sieuwerts AM, Luider TM, van der Weiden M, Sillevis-Smitt PA, Kros JM. Differential expression of splicing variants of the human caldesmon gene (CALD1) in glioma neovascularization versus normal brain microvasculature. The American journal of pathology. 2004;164(6):2217-28.
25. Kaneko M, Iwase I, Yamasaki Y, Takai T, Wu Y, Kanemoto S, et al. Genome-wide identification and gene expression profiling of ubiquitin ligases for endoplasmic reticulum protein degradation. Scientific reports. 2016;6:30955.
26. Kim SJ, Lee SY, Lee C, Kim I, An HJ, Kim JY, et al. Differential expression profiling of genes in a complete hydatidiform mole using cDNA microarray analysis. Gynecologic oncology. 2006;103(2):654-60.
27. Du J, Liu X, Wu Y, Zhu J, Tang Y. Essential role of STX6 in esophageal squamous cell carcinoma growth and migration. Biochemical and biophysical research communications. 2016;472(1):60-7.
28. Hoang TT, Sikdar S, Xu C-J, Lee MK, Cardwell J, Forno E, et al. Epigenome-Wide Association Study of DNA Methylation and Adult Asthma in the Agricultural Lung Health Study. European Respiratory Journal. 2020.
29. Bahr TM, Hughes GJ, Armstrong M, Reisdorph R, Coldren CD, Edwards MG, et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. American journal of respiratory cell and molecular biology. 2013;49(2):316-23.
30. Qiao D, Lange C, Beaty TH, Crapo JD, Barnes KC, Bamshad M, et al. Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2016;193(12):1353-63.
31. Vonk J, Nieuwenhuis M, Dijk F, Boudier A, Siroux V, Bouzigon E, et al. Novel genes and insights in complete asthma remission: A genome-wide association study on clinical and complete asthma remission. Clinical & Experimental Allergy. 2018;48(10):1286-96.
32. Yin W, Chen J, Wang G, Zhang D. MicroRNA-106b functions as an oncogene and regulates tumor viability and metastasis by targeting LARP4B in prostate cancer. Molecular Medicine Reports. 2019;20(2):951-8.
33. Martın-Martın B, Nabokina SM, Blasi J, Lazo PA, Mollinedo F. Involvement of SNAP-23 and syntaxin 6 in human neutrophil exocytosis. Blood, The Journal of the American Society of Hematology. 2000;96(7):2574-83.
34. VK V. Chronic obstructive pulmonary disease. Indian J Med Res. 2013;137(2):251-69.
35. Barnes PJ, Shapiro SD, Pauwels R. Chronic obstructive pulmonary disease: molecular and cellularmechanisms. European Respiratory Journal. 2003;22(4):672-88.
36. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.

37. Mostafaei S, Kazemnejad A, Jamalkandi SA, Amirhashchi S, Donnelly SC, Armstrong ME, et al. Identification of Novel Genes in Human Airway Epithelial Cells associated with Chronic Obstructive Pulmonary Disease (COPD) using Machine-Based Learning Algorithms. Scientific reports. 2018;8(1):1-20.
38. Zhao P, Yu B. On model selection consistency of Lasso. Journal of Machine learning research. 2006;7(Nov):2541-63.
39. Wang L, Kim Y, Li R. Calibrating non-convex penalized regression in ultra-high dimension. Annals of statistics. 2013;41(5):2505.
40. Singh D, Fox SM, Tal-Singer R, Plumb J, Bates S, Broad P, et al. Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. Thorax. 2011;66(6):489-95.
41. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3.
42. Wilson CL, Miller CJ. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. Bioinformatics. 2005;21(18):3683-5.
43. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: Springer; 2005. p. 397-420.
44. Ihaka R, Gentleman R. R: a language for data analysis and graphics. Journal of computational and graphical statistics. 1996;5(3):299-314.
45. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association. 2001;96(456):1348-60.
46. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics. 2004;32(3):928-61.
47. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics. 2010;38(2):894-942.
48. Breheny P, Breheny MP. Package 'ncvreg'. 2020.
49. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence. 2010;32(3):569-75.
50. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. The Lancet. 2003;362(9381):362-9.
51. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5):352-9.
52. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12(1):1-8.
53. Bacardit J, Llorà X. Large-scale data mining using genetics-based machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2013;3(1):37-61.
54. Galili T. heatmaply: interactive heat maps (with R). Month. 2016;545.

Figure legend:

Figure 1: The ROC curves for logist regressions using LASSO, SCAD, and MCP regularization.

Figure 2: Spearman's rank correlation, co-expression matrix between the selected genes: heatmap for hierarchical clustering the twenty-four candidate genes based on their pattern of gene expression.

Figure 3: Overlapping between LASSO, SCAD, and MCP regularized logistic regressions for genes selection

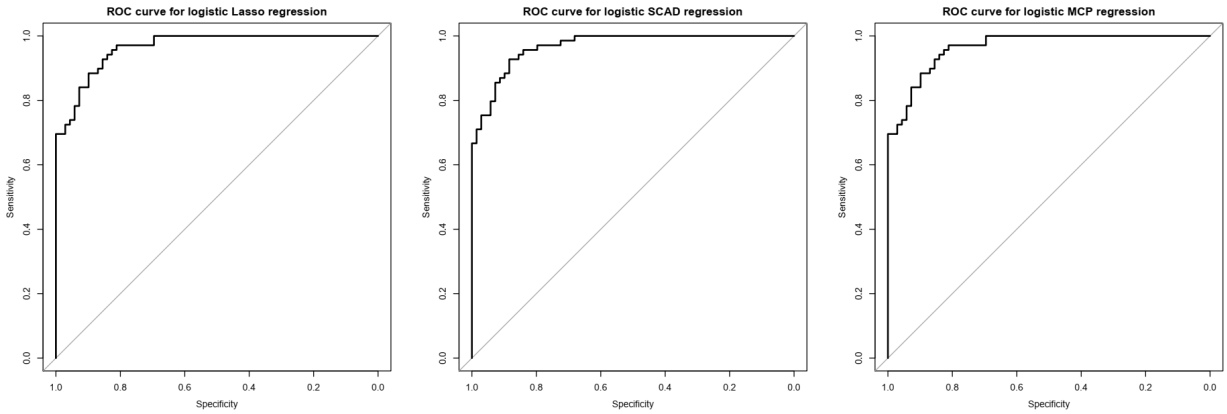

Figure 1: The ROC curves for logist regressions using LASSO, SCAD, and MCP regularization.

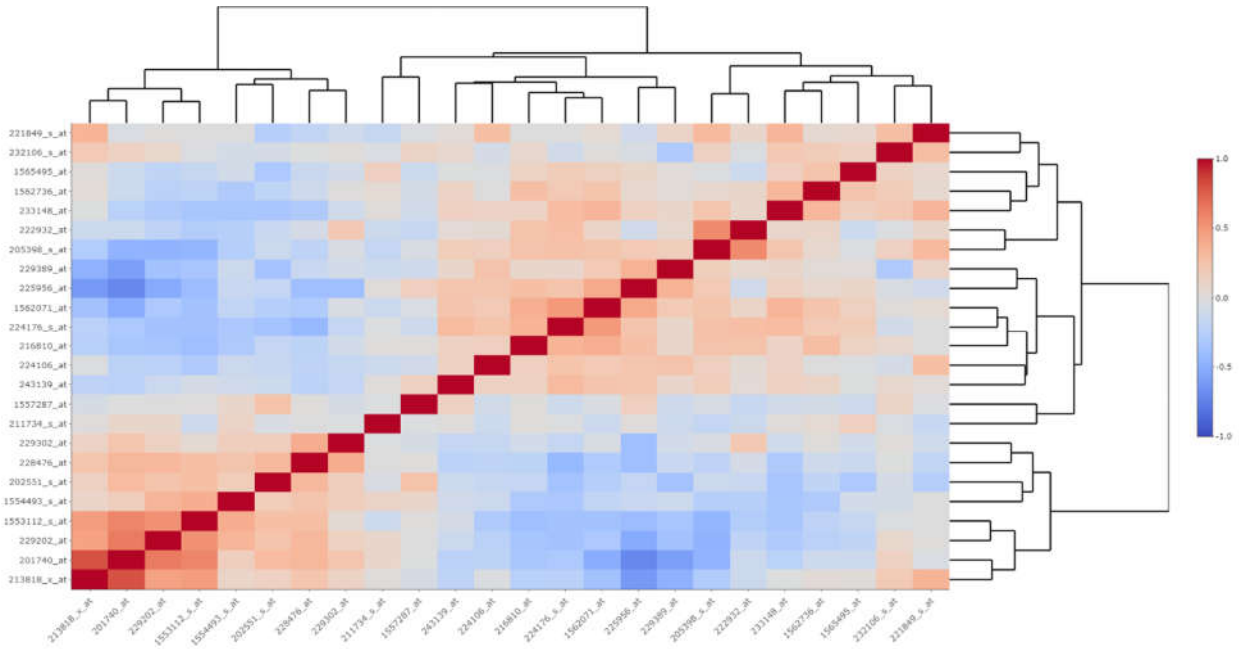

Figure 2: Spearman's rank correlation, co-expression matrix between the selected genes: heatmap for hierarchical clustering the twenty-four candidate genes based on their pattern of gene expression.

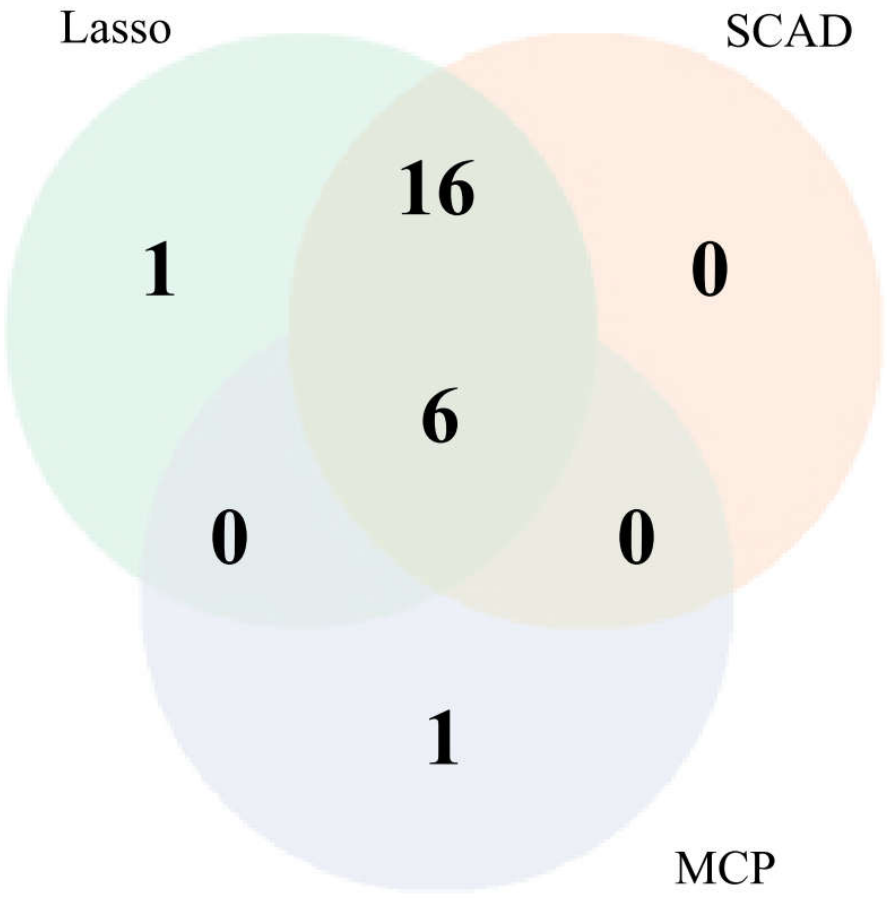

Figure 3: Overlapping between 3 methods of logistic regression for genes selection

Table 1: Results of gene selection by Lasso, SCAD and MCP logistic regression

Gene symbol	Gene title	Gene ID	LASSO				SCAD				MCP			
			Estimate	z	P-value	OR	Estimate	z	P-value	OR	Estimate	z	P-value	OR
LOC100507634	uncharacterized LOC100507634	1557566_at	0.016	2.652	0.012	1.016	0.011	2.692	0.011	1.011				
		1559495_at	0.015	2.759	0.009	1.015	0.016	2.823	0.007	1.016				
LINC00693	long intergenic non-protein coding RNA 693	1564250_at	-0.073	-3.042	0.004	0.93	-0.063	-3.055	0.004	0.939				
LOC102724382	uncharacterized LOC102724382	1568887_at	0.024	2.881	0.006	1.024	0.022	2.934	0.005	1.022				
TAF15	TATA-box binding protein associated factor 15	202840_at	-0.046	-2.996	0.004	0.955	-0.036	-3.007	0.004	0.965				
CACNA1G	calcium voltage-gated channel subunit alpha1 G	207869_s_at	0.025	2.671	0.011	1.025	0.021	2.711	0.01	1.021				
RPIA	ribose 5-phosphate isomerase A	212973_at	-0.189	-3.685	0	0.828	-0.18	-3.712	0	0.836				
PLCB1	phospholipase C beta 1	213222_at	0.08	2.998	0.004	1.083	0.075	3.04	0.004	1.078				
PELP1	proline, glutamate and leucine rich protein 1	215354_s_at	-0.108	-3.399	0.001	0.898	-0.114	-3.495	0.001	0.892	-0.517	-5.33	0	0.6
SLC38A2	solute carrier family 38 member 2	222982_x_at	0.077	3.163	0.003	1.081	0.083	3.252	0.002	1.087				
ZNF33A	zinc finger protein 33A	224276_at	0.026	2.87	0.006	1.027	0.026	2.932	0.005	1.027				
ESYT2	extended synaptotagmin 2	224699_s_at	0.026	2.764	0.009	1.027	0.022	2.805	0.008	1.023				
STIM2	stromal interaction molecule 2	225250_at	-0.195	-2.411	0.021	0.823								
LAMA1	laminin subunit alpha 1	227048_at	-0.158	-3.526	0.001	0.854	-0.152	-3.563	0.001	0.859	-0.036	-3.46	0.001	0.97
RNF130	ring finger protein 130	230932_at	0.171	3.596	0.001	1.186	0.173	3.674	0	1.188	0.193	4.031	0	1.21
LOC100288675	uncharacterized LOC100288675	234344_at	0.012	2.772	0.009	1.012	0.008	2.815	0.008	1.008				
AMOTL1	angiomotin like 1	235277_at	-0.216	-3.869	0	0.806	-0.218	-3.952	0	0.804	-0.393	-4.86	0	0.68
TMEM182	transmembrane protein 182	238578_at	-0.106	-3.295	0.002	0.9	-0.102	-3.341	0.002	0.903				
LARP4B	La ribonucleoprotein domain family member 4B	240005_at	0.016	2.697	0.011	1.017	0.012	2.736	0.009	1.012				
		241300_at	-0.029	-2.705	0.01	0.971	-0.023	-2.731	0.01	0.978				
STX6	syntaxin 6	244041_at	0.083	3.051	0.004	1.087	0.078	3.091	0.003	1.081	0.149	3.741	0	1.16
		244418_at	0.138	3.437	0.001	1.147	0.135	3.489	0.001	1.144				
PRDX2	peroxiredoxin 2	39729_at	-0.126	-3.425	0.001	0.882	-0.122	-3.474	0.001	0.885	-0.037	-3.5	0.001	0.96
CALD1	caldesmon 1	212077_at									-0.015	-3.31	0.002	0.99

Table 1 should be palced in page 6.