Prediction of Psoriasis from Gene Expression Profiling Using Penalized Logistic Regression Model

Payam Amini[1] iD **, Leili Tapak[2,3]** iD **, Saeid Afshar[4]** iD **, Mahlagha Afrasiabi[5]** iD **,**

MohammadKazem Ghasemi[6] iD **, Pedram Alirezaei[7]** iD

1. *School of Medicine, Keele University, Keele, Staffordshire, ST5 5BG, The United Kingdom*
2. *Dept. of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
3. *Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran*
4. *Dept. of Medical Biotechnology, School of Advanced Medical Sciences and Technologies, Hamadan University of Medical Sciences, Hamadan, Iran*
5. *Dept. of Computer, Hamedan University of Technology, Hamedan, Iran*
6. *Student Research Committee, Hamadan University of Medical Sciences, Hamadan, Iran*
7. *Psoriasis Research Center, Hamadan University of Medical Sciences, Hamadan, Iran*

Article Info

Corresponding Information:
Leili Tapak,
Dept. of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

E-Mail: l.tapak06@gmail.com

ABSTRACT

Background & Objective: Psoriasis is one of the most common skin disorders in humans and is believed to have genetic foundations. The aim of this study is to identify potential genetic biomarkers for psoriasis using penalized methods.

Materials & Methods: The gene chip GSE55201, which included 74 individuals (34 patients with psoriasis and 30 healthy individuals), was obtained from GEO. Three penalized approaches were used in logistic regression, including Least Absolute Shrinkage Selection Operator, Minimax Concave Penalty, and Smoothing Clipped Absolute Deviation, to identify the most important genes associated with psoriasis. To validate the results, Random Forest was used to assess the predictive power of the selected genes in a validation dataset.

Results: The analysis identified ADORA3 and C16orf72 as two genes that were commonly associated with psoriasis. The independent samples t-test revealed significantly higher expression of ADORA3 and C16orf72 among psoriasis cases (p<0.001). The area under the ROC curve for predicting psoriasis was 0.88 (95% CI: 0.80-0.96) for ADORA3 and 0.75 (95% CI: 0.75-0.94) for C16orf72. The Random Forest analysis showed that the model using these genes had a prediction probability of 0.68 (95% CI: 0.53-0.83).

Conclusion: Among all the methods used, MCP outperformed other penalties, selecting a smaller subset with compatible performance. Two key genes, ADORA3 and C16orf72, were found to be associated with psoriasis and were identified for further study. These genes may serve as genetic biomarkers for predicting psoriasis.

Keywords: Gene Expression Profiling, Bioinformatics, Psoriasis, Biomarker, Prognosis, Penalized Regression

Introduction

Psoriasis is a hereditary autoimmune skin condition that is characterized by abnormally enlarged skin patches **(1)**. The red and scaly plaques that characterize this chronic skin condition are most commonly found on the elbows, knees, scalp, and lower back **(2)**. Psoriasis has significant psychological and social consequences, as well as imposing a high financial burden on both sufferers and healthcare systems **(3)**. It has been suggested that many countries lack epidemiological data on this immune-mediated inflammatory skin condition. Although psoriasis is more common in adults than in children, studies have shown that it can affect people of all ages, particularly those between the ages of 20-30 and 50-60

(4). Psoriasis cases have been documented in greater numbers in developed economies **(5)**. Psoriasis affects approximately 2 to 4% of the population in Western countries **(6)**.

Although the origins of this non-communicable illness are unknown, several recognized risk factors include infections and skin injuries, family history, weather, certain medications, stress, smoking and exposure to secondhand smoke, obesity, and alcohol use. Psoriasis can cause severe symptoms such as burning, stinging, itching, discomfort, and skin cracking in some cases. The primary goal of psoriasis treatment is to limit cell proliferation. Psoriasis is classified into three stages: mild, moderate,

and severe. Topical medications, such as calcineurin inhibitors and keratolytics, topical corticosteroids, and vitamin D analogues, can be used to treat moderate cases (7). Individuals with moderate to severe psoriasis may require more aggressive treatments, such as systemic medications, UV radiation, and rotational therapy, which involves switching medications to reduce the risk of systemic treatment toxicity over time (8). Although molecular tissue research on psoriasis is still ongoing, its genetic basis has been established. Quantitative polymerase chain reaction (qPCR) and microarray approaches have been used in comprehensive screening to identify key genes and the regulator of stem cell proliferation involved in the development of psoriasis, as well as to uncover patterns of gene expression) .4) Psoriasis lesions are significantly associated with immune cells such as CD3+ T cells and CD11c+ dendritic cells, and the cytokines produced by these cells. In particular, tumor necrosis factor-alpha (TNFα), interferon-gamma (IFNγ), interleukin-17 (IL-17), IL-22, IL-23, IL-12, and IL-1beta are linked to the pathogenesis of psoriasis through the activation of keratinocytes and other skin cells in the dermis (8).

The selection of the most important genes from high-dimensional and correlated data is a crucial step. To assess the impact of the most important genes among a large number of genes, researchers can reduce the dimensionality of the data by using important components or employing variable selection approaches. While it has been suggested to reduce the dimensionality of data using principal component analysis and partial least squares approaches, these methods may lose some information from the data and can be difficult to interpret (9). Penalized logistic regression is a useful method for feature selection, allowing for simultaneous estimation and gene selection (10).

In this study, we employed three approaches in penalized logistic regression, including Least Absolute Shrinkage Selection Operator (LASSO), Minimax Concave Penalty (MCP), and Smoothing Clipped Absolute Deviation (SCAD), to select the most important genes associated with psoriasis.

Materials and Methods

Data Source and Preprocessing

We used the publicly available dataset of psoriasis whole blood transcriptome dataset (available in GEO repository: GSE55201 dataset generated using the Affymetrix U133 Plus (microarray) with platform ID GPL570). The information provided in this dataset is the gene expression of 30 healthy controls and 44 psoriasis patients at baseline and seven psoriasis patients after two weeks of treatment (11). In this study, the data of 30 healthy controls and 44 psoriasis patients at baseline samples were used.

Ethical approval

This study was approved by the Ethics Committee of Hamadan University of Medical Sciences (approval code: IR.UMSHA.REC.1398.637). All methods were performed in accordance with the relevant guidelines and regulations.

Penalized logistic regression

To classify cases as psoriasis or non-psoriasis, three penalized logistic regression models with LASSO, MCP, and SCAD penalty functions were employed to identify important features. The logistic regression approach was selected for its ability to directly estimate the effect size and lack of assumptions about predictor distribution. Let $\mathbf{Y} = (y_1,...,y_n)$ be the vector of "n" binary outcome variables, X_{ij} is an $n \times p$ matrix of "p" gene expression for cases, and the β_i be a "p"-sized vector of the regression coefficients indicating the measure of effect size. Then the model is written as follows:

$$log(\frac{p(y_1 = 1)}{p(y_1 = 0)}) = \sum_{i=1}^{p} \beta_i X_i$$

The LASSO approach forces the coefficients of some less contributive variables to be exactly zero, retaining the most significant genes in the final model by conditioning $\sum_{i=1}^{k} |\beta_i| \le \lambda$ with λ as the tuning parameter. This method penalizes the model using a penalty term, the sum of the absolute regression coefficients. LASSO performs well when some predictors have large effect sizes, and the remaining predictors have very small coefficients. In the case of highly pairwise correlated genes, this method selects only one gene.

The SCAD method selects genes and estimates their effect size simultaneously. It retains the sparsity feature of LASSO by implementing another penalty to small parameters, a constant penalty for large parameters to cope with bias, and a quadratic spline to generate a continuous differentiable penalty function. Despite its popular properties, SCAD is non-convex, causing difficulties in computation. The penalty function for SCAD is as follows:

$$Penalty(|\beta|) = \begin{cases} \lambda|\beta| & 0 \le |\beta| \le \lambda \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & \lambda \le |\beta| \le a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & |\beta| \ge a\lambda \end{cases}$$

The MCP is similar to SCAD and covers sparsity and unbiasedness. It is shown as:

$$Penalty(|\beta|) = \begin{cases} \lambda\left(|\beta| - \dfrac{|\beta|^2}{2\lambda a}\right) & |\beta| \le a\lambda \\ \dfrac{a\lambda^2}{2} & |\beta| \ge a\lambda \end{cases}$$

It minimizes the maximum concavity using a tuning parameter to achieve the least unbiasedness and highest

concavity. Similar to SCAD, computational problems are still an issue with MCP.

In the present study, the penalized logistic regression model with LASSO, SCAD and MCP penalties was used to select gene expression profiles related to the binary response variable of having psoriasis, and not having psoriasis (healthy control). There were 54,657 original probe sets (i.e., *p*) compared to the sample size of 74 (30 healthy controls and 44 psoriasis patients). so, *p* was much larger than *n* and a sparse subset of gene profiles was selected using the penalized logistic model by using all 54,657 original probe sets as inputs of the model. In this method, there were two parameters, λ and a, that should be optimized. The optimal values of these parameters were selected through cross-validation, as it is explained in the following section.

Tuning Parameters

In this study, the optimal value of each tuning parameter was determined using a 10-fold cross-validation strategy. The data were divided into training and testing sets (70-30), with the testing set left out for external validation of the methods. The training data set was randomly split into ten subsets. The penalized models were then fitted ten times, with one of the ten subsets left out as the testing set each time and the remaining nine subsets used as the training set. The models were implemented for a range of values for the parameter λ, starting from zero (no shrinkage) through a value that puts maximum shrinkage on the regression parameters. The optimal λ was chosen based on the smallest Bayesian Information Criterion (BIC) over the testing sets

Evaluation Criteria for External validation

In this study, the total accuracy and the area under the ROC curve (AUC) were used to evaluate the performance of the models.

Software

Data analysis was carried out utilizing R software (version 4.1.1) and "Glmnet" package.

Results

The penalized logistic regression using three penalty functions identified different subsets of genes: the MCP selected three genes, the SCAD identified 15, and the LASSO approach selected 24 genes. Among the selected genes, ADORA3 and C16orf72 were common to all three penalized methods. The Bayesian Information Criterion (BIC) values for logistic regression models based on each subset of selected genes were 56, 28, and 58 for SCAD, MCP, and LASSO, respectively, indicating that the MCP selected the best subset of genes. The area under the receiver operating characteristic (ROC) curve was 1.000, 0.967, and 1.000 for SCAD, MCP, and LASSO, respectively. This suggests that the MCP could select a better subset of genes (of size 3), as its classification performance in distinguishing psoriasis patients from healthy controls was comparable to that of SCAD and LASSO with larger numbers of selected genes (15 and 24, respectively). Table 1 shows the names of the unique genes resulting from each approach, and Figure 1 displays a Venn diagram of the identified genes by the three penalized approaches.

Table 1. The resulted genes found by the three penalized approaches

Penalized methods		
SCAD	**MCP**	**LASSO**
ADORA3, C16orf72, IL5RA, ARHGAP26, RPS15A, ABHD2, ABHD2, TOP2A, ASPH, SIGIRR, JPX, FTSJ3, MAPRE1, RBM25, PPAN-P2RY11	*ADORA3, C16orf72, CRCP*	*ADORA3, C16orf72, CRCP, RPS15A, ABHD2, ABHD2, TOP2A, USP9Y, INTS7, ASPH, NRG1, JPX, PCLAF, MT-ND6, ZNRF1, MAPRE1, SEC22B, NRG1, ANP32A, LUZP1, VHL, JAKMIP2, ATG2B, TOP3A*

LASSO: Least Absolute Shrinkage Selection Operator, MCP: Minimax Concave Penalty, SCAD: Smoothing Clipped Absolute Deviation
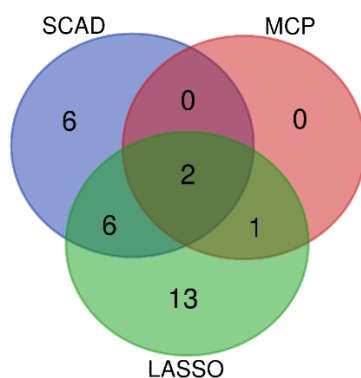

Figure 1. Venn diagram of identified genes by the three penalized approaches

The results of the independent samples t-test are presented in Table 2, which shows that the mean expression levels of ADORA3 and C16orf72 are significantly higher in psoriasis cases than in non-psoriasis cases (p<0.001). Furthermore, ADORA3 and C16orf72 were found to be predictors of psoriasis, with areas under the receiver operating characteristic (ROC) curve of 0.88 (95% CI: 0.80-0.96) and 0.75 (95% CI: 0.75-0.94), respectively. The ROC curve for this model is shown in Figure 2.

Table 2. Results of the independent samples t-test and the area under the ROC curve for *ADORA3* and C16orf72 expressions

Gene	Non-Psoriasis Mean (SD)	Psoriasis Mean (SD)	T value	P-value	AUC (95% CI)
ADORA3	6.279 (0.955)	7.819 (0.858)	7.091	<0.001	0.88(0.80-0.96)
C16orf72	8.931(0.324)	9.367(0.308)	5.806	<0.001	0.75(0.75-0.94)

SD: Standard Deviation; AUC: The area under the ROC curve; CI: Confidence Interval
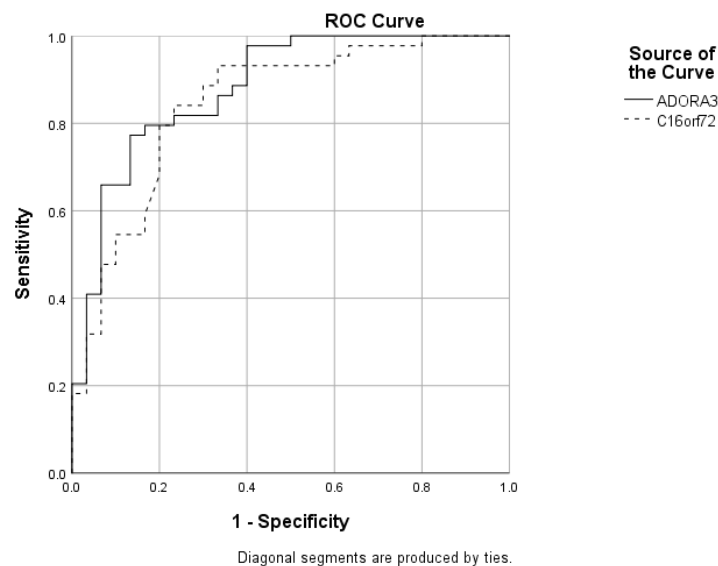
Figure 2. ROC curve analysis of predicting psoriasis using *ADORA3* and C16orf72

Random forest was utilized to determine the predictive power of selected genes in the validation set of data. The area under the ROC curve (95% confidence interval) after applying the random forest approach was 0.68 (0.53-0.83) which showed a significant prediction by the genes in the model (p=0.034). Figure 3 demonstrates ROC curve predicting psoriasis using ADORA3 and C16orf72 in the validation dataset resulted from the random forest.
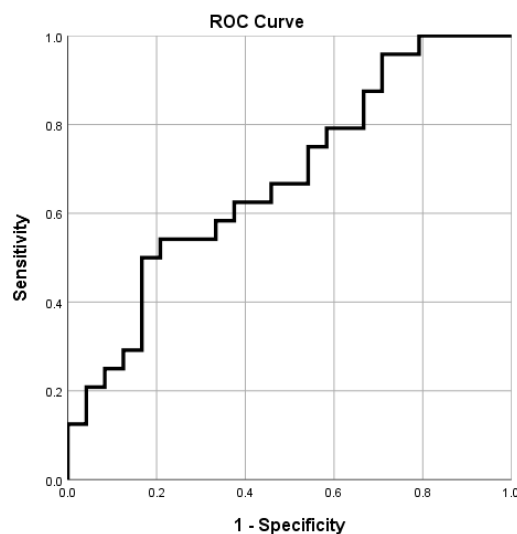
Figure 3. Random forest ROC curve predicting Psoriasis using ADORA3 and C16orf72 in the validation dataset

Discussion

Psoriasis is a chronic autoimmune/inflammatory disease with a genetic basis that affects multiple aspects of patients' health. It is associated with increased risks of various diseases, including hyperlipidemia, hypertension, coronary artery disease (CAD), diabetes (type II), stroke, and myocardial infarction. Identifying prognostic markers using state-of-the-art models is critical for early detection and treatment of psoriasis. In this study, we applied three penalizing functions in logistic regression to classify cases as psoriasis or non-psoriasis based on potential features. All the methods identified ADORA3 and C16orf72 as highly associated with the presence of psoriasis among many genes. Using a validation dataset, we applied random forest, a classification approach, and found that the selected two genes are significant predictors of psoriasis.

The protein encoded by the adenosine A3 receptor gene (ADORA3) is a G-protein-coupled receptor that is involved in the regulation of several biological functions through the inhibition of adenylate cyclase activity [12]. ADORA3 has an essential role in the inhibition of inflammation [13]. Our findings indicate that the expression level of ADORA3 is overexpressed in psoriatic samples, which is consistent with the results of previous studies. Stamp et al. reported a significant increase in the expression level of ADORA3 in synovial samples of patients with rheumatoid arthritis, another autoimmune disease [14]. Studies have also shown that ADORA3 is present in peripheral blood mononuclear cells of patients with autoimmune inflammatory diseases, including rheumatoid arthritis [15] psoriasis, and Crohn's disease [16]. They also reported a significant reduction of IL-1b/IL-6 release in RA patients mediated by the ADORA2a and ADORA3 stimulation. Moreover, variations in the expression of ADORA subtypes have been observed and attributed to the tissues affected by the disease [17].

C16orf72, also known as Telomere Attrition and P53 Response 1 (TAPR1), plays a critical role in the negative regulation of P53 signaling induced by the loss of telomere integrity [18]. This gene "exhibits a strong synthetic genetic interaction with telomerase inhibition or deletion of TERT, and appears to taper the response to p53 upon loss of telomere integrity" 19. Inflammation accelerates cellular senescence due to the shortening of telomere length, and recent studies have indicated that telomere length plays a role in the pathogenesis of various autoimmune diseases [19, 20].

As technology has developed, genetics research has produced increasingly high-dimensional data that can be analyzed to gain insights into disease mechanisms. To avoid assessing irrelevant genes and obtain a subset of genes with high classification capabilities, dimension reduction approaches such as penalized logistic regression can be used. However, different penalized methods have their own advantages and disadvantages. LASSO is feasible for high-dimensional computations but has limitations in statistical accuracy for prediction, the number of selected genes due to the number of samples, and equality in penalizing all gene coefficients [21]. SCAD provides a good goodness of fit and promotes a sparse estimate of the regression [22]. In contrast, MCP utilizes similar rates of penalization to Lasso and decreases the rate to zero as the absolute value of the coefficient increases [23]. In our study, these penalized logistic regression models identified the most associated genes with the presence of psoriasis.

Conclusion

Our findings showed that MCP outperformed other penalties by selecting a smaller subset with compatible performance. This study identified a small subset of genes, including ADORA3 and TAPR1, that can be used for the prognosis of psoriasis. Further laboratory investigations are needed to evaluate the identified genes.

Acknowledgments

Data Availability

A publicly available dataset was analyzed in this study. This data can be found at https://www.ncbi.nlm.nih.gov/geo/ (the NCBI Gene Expression Omnibus).

Funding

Conflict of Interest

The authors declare that there is no conflict of interests.

Authors' Contribution

Payam Amini: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing-original draft, writing.

Leili Tapak: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing-original draft, writing-review and editing, visualization, supervision, project administration, funding acquisition.

Saeid Afshar: Investigation, data curation, review and editing,

Mahlagha Afrasiabi: Investigation, data curation, software, review and editing.

MohammadKazem Ghasemi: Formal analysis, writing-review and editing.

Pedram Alirezaei: Conceptualization, methodology, validation, writing-review and editing.

References

1. Rendon A, Schäkel K. Psoriasis pathogenesis and treatment. Int J Molec Sci. 2019; 20(6):1475. [DOI:10.3390/ijms20061475] [PMID] [PMCID]

2. Stern RS, Nijsten T, Feldman SR,et al. Psoriasis is common, carries a substantial burden even when not extensive, and is associated with widespread treatment dissatisfaction. J Investig Dermatol Symp Proc. 2004;9(2):136-9.[PMID] [DOI:10.1046/j.1087-0024.2003.09102.x]

3. Javitz HS, Ward MM, Farber E, Nail L, Vallow SG. The direct cost of care for psoriasis and psoriatic arthritis in the United States. J Am Acad Dermatol. 2002; 46(6):850-60. [DOI:10.1067/mjd.2002.119669] [PMID]

4. Tapak L, Afshar S, Afrasiabi M, Ghasemi MK, Alirezaei P. Application of genetic algorithm-based support vector machine in identification of gene expression signatures for psoriasis classification: a hybrid model. Bio Med Res Int. 2021;2021:55207. [DOI:10.1155/2021/5520710] [PMID] [PMCID]

5. Parisi R, Iskandar IY, Kontopantelis E, Augustin M, Griffiths CE, Ashcroft DM. National, regional, and worldwide epidemiology of psoriasis: systematic analysis and modelling study. BMJ. 2020;369. [DOI:10.1136/bmj.m1590] [PMID] [PMCID]

6. Parisi R, Symmons DP, Griffiths CE, Ashcroft DM. Global epidemiology of psoriasis: a systematic review of incidence and prevalence. J Investig Dermatol. 2013;133(2):377-85. [DOI:10.1038/jid.2012.339] [PMID]

7. Armstrong AW, Read C. Pathophysiology, clinical presentation, and treatment of psoriasis: a review. JAMA. 2020;323(19):1945-60. [DOI:10.1001/jama.2020.4006] [PMID]

8. Johnson-Huang LM, Lowes MA, Krueger JG. Putting together the psoriasis puzzle: an update on developing targeted therapies. Dis Model Mech. 2012;5(4):423-33. [DOI:10.1242/dmm.009092] [PMID] [PMCID]

9. Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction. J Soft Comput Data Min. 2021;2(1):20-30. [DOI:10.30880/jscdm.2021.02.01.003]

10. Zhu Y, Tan TL, Cheang WK. Penalized logistic regression for classification and feature selection with its application to detection of two official species of Ganoderma. Chemometr Intell Lab Syst. 2017;171:55-64. [DOI:10.1016/j.chemolab.2017.09.019]

11. Wang CQ, Suarez-Farinas M, Nograles KE, et al. IL-17 induces inflammation-associated gene products in blood monocytes, and treatment with ixekizumab reduces their expression in psoriasis patient blood. J Investig Dermatol. 2014;134(12): 2990. [DOI:10.1038/jid.2014.268] [PMID]

12. Zhao Z, Ravid S, Ravid K. Chromosomal mapping of the mouse A3 adenosine receptor gene,adora3. Genomics. 1995;30(1):118-9. [DOI:10.1006/geno.1995.0023] [PMID]

13. Roszkiewicz J, Michałek D, Ryk A, Swacha Z, Szmyd B, Smolewska E. The impact of single nucleotide polymorphisms in ADORA2A and ADORA3 genes on the early response to methotrexate and presence of therapy side effects in children with juvenile idiopathic arthritis: Results of a preliminary study. Int J Rheum Dis. 2020;23(11):1505-13. [PMID] [DOI:10.1111/1756-185X.13972]

14. Stamp LK, Hazlett J, Roberts RL, Frampton C, Highton J, Hessian PA. Adenosine receptor expression in rheumatoid synovium: a basis for methotrexate action. Arthritis Res Ther. 2012; 14(3):R138-R. [DOI:10.1186/ar3871] [PMID] [PMCID]

15. Varani K, Padovan M, Vincenzi F, et al. A2A and A3 adenosine receptor expression in rheumatoid arthritis: upregulation, inverse correlation with disease activity score and suppression of inflammatory cytokine and metalloproteinase release. Arthritis Res Ther. 2011;13(6):1-13. [DOI:10.1186/ar3527] [PMID] [PMCID]

16. Ochaion A, Bar-Yehuda S, Cohen S, et al. The anti-inflammatory target A3 adenosine receptor is over-expressed in rheumatoid arthritis, psoriasis and Crohn's disease. Cell Immunol. 2009;258(2): 115-22. [DOI:10.1016/j.cellimm.2009.03.020] [PMID]

17. Stamp LK, Hazlett J, Roberts RL, Frampton C, Highton J, Hessian PA. Adenosine receptor expression in rheumatoid synovium: a basis for methotrexate action. Arthritis Res Ther. 2012; 14(3):1-9. [DOI:10.1186/ar3871] [PMID] [PMCID]

18. Benslimane Y, Sánchez-Osuna M, Coulombe-Huntington J, et al. A novel p53 regulator, C16ORF72/TAPR1, buffers against telomerase

inhibition. Aging Cell. 2021;20(4):e13331-e. [DOI:10.1111/acel.13331] [PMID] [PMCID]

19. Hohensinner PJ, Goronzy JJ, Weyand CM. Telomere dysfunction, autoimmunity and aging. Aging Dis. 2011;2(6):524-37.

20. Georgin-Lavialle S, Aouba A, Mouthon L, et al. The telomere/telomerase system in autoimmune and systemic immune-mediated diseases. Autoimmun Rev. 2010;9(10):646-51. [DOI:10.1016/j.autrev.2010.04.004] [PMID]

21. Fan J, Fan Y, Barut E. Adaptive robust variable selection. Ann Statis. 2014;42(1):324. [DOI:10.1214/13-AOS1191] [PMID] [PMCID]

22. Pang S, Havukkala I, Hu Y, Kasabov N. Classification consistency analysis for bootstrapping gene selection. Neural Comput Applic. 2007;16(6):527-39. [DOI:10.1007/s00521-007-0110-1]

23. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statis Comput. 2015;25(2):173-87. [PMID] [PMCID] [DOI:10.1007/s11222-013-9424-2]

How to Cite This Article:

Amini P, Tapak L, Afshar S, Afrasiabi M, Ghasemi M, Alirezaei P. Prediction of Psoriasis from Gene Expression Profiling Using Penalized Logistic Regression Model. J Adv Med Biomed Res. 2023; 31(138):481-7.

Download citation:
BibTeX | **RIS** | **EndNote** | **Medlars** | **ProCite** | **Reference Manager** | **RefWorks**

Send citation to:
Mendeley   **Zotero**   **RefWorks**