# TRANSCRIPTOMICS AND THE "CURSE OF DIMENSIONALITY": MONTE CARLO SIMULATIONS OF ML-MODELS AS A TOOL FOR ANALYZING MULTIDIMENSIONAL DATA IN TASKS OF SEARCHING MARKERS OF BIOLOGICAL PROCESSES

**© 2025 G. Zh. Osmak [a, b, *], M. V. Pisklova [a, b]**

[a] *National Medical Research Center of Cardiology named after ac. E.I. Chazov of the Ministry of Health of the Russian Federation, Moscow, 121552 Russia*

[b] *Pirogov Russian National Research Medical University of the Ministry of Health of the Russian Federation, Moscow, 117997 Russia*

**e-mail: german.osmak@gmail.com*

Received April 11, 2024

Revised May 06, 2024

Accepted May 26, 2024

**Abstract.** High-throughput transcriptomic research methods provide the assessment of a vast number of factors, valuable for researchers. At the same time the "curse of dimensionality" issues arise, which lead to increasing requirements on data processing and analysis methods. In this study, we propose a new algorithm that combines Monte Carlo methods and machine learning. This algorithm will enable feature space reduction by highlighting genes most likely associated with the investigated diseases. Our approach allows not only to generate a set of "interesting" genes but also to assign weight to each gene, indicating its "importance". This measure can be used in subsequent statistical analysis, visualization, and interpretation of results. Algorithm performance was demonstrated on open transcriptomic data of patients with HCM (GSE36961 and GSE1145). The analysis revealed genes *MYH6, FCN3, RASD1*, and *SERPINA3*, which is in good agreement with the available literature.

**Keywords**: *transcriptomics, machine learning, Monte Carlo, hypertrophic cardiomyopathy, biomarkers*

**DOI:** 10.31857/S00268984250111e4

# INTRODUCTION

Transcriptomics and high-throughput research methods such as RNA sequencing (RNA-seq) or microarrays (MicroArray) today undoubtedly occupy an important place in the arsenal of tools for studying the molecular mechanisms of biological systems, the pathogenesis of various diseases and the search for their markers.[1].

Identification of differentially expressed genes (DEGs) or transcripts under different conditions (comparison groups) is one of the important tasks of transcriptome profiling. Differential expression data are usually presented in a matrix format, where each row corresponds to a gene (or transcript), and each column corresponds to a sample, with the cells indicating the gene expression level in the sample.[2]. The main research problem is to detect statistically significant DEGs between different groups of samples (e.g. healthy and sick). One of the frequent problems that arise in statistical processing of such data is related to the "curse of dimensionality"[3].

The "curse of dimensionality" is a phenomenon in which the feature space increases with increasing number of dimensions or input variables, which can lead to increased noise and erroneous conclusions. The average feature space dimensionality of transcriptome profiling data is over 10,000. The average sample size is less than 100 points. Thus, despite the richness of information obtained by high-throughput methods, interpretation of these data can be challenging due to the large number of genes and small number of samples.

Standard means of solving the indicated problem include various tools for adjusting values*p*-value with multiple comparisons, widely used inside popular packages like EdgeR[4]or Limma[5].

In this paper, we propose a new approach based on machine learning (ML) methods for reducing data dimensionality and identifying key genes with the highest chance of being associated with the studied disease, followed by the application of weighted correction procedures for multiple comparisons. The weights for adjusting $p$ -values are also obtained using ML methods.

The essence of the approach is to use Monte Carlo simulations to generate classifiers with high generalization ability on transcriptome profiling data. Then, features important for their operation, or key genes, are extracted from these classifiers, and a reduced feature space is formed for subsequent testing of association hypotheses using standard methods. The resulting feature space will also be a weighted space, i.e., with a weight function or measure defined on it. The weight will be defined as the proportion of models in which the gene was included, multiplied by the ROC-AUC quality metric, averaged across these models. This weight will be used when conducting weighted correction

procedures for multiple hypothesis testing, such as weighted Bonferroni, Holm, or Benjamini-Hochberg methods.

Initially, the listed weighted correction methods were developed to account for prior information [6, 7] . Currently, most of the work devoted to the development of these methods reduces to the problem of maximizing the power of statistical tests by the weight vector [8, 9] . In the presented work, we propose to return to the classical formulation with the assignment of weight coefficients reflecting some prior information, which we obtain from the data ( *data driven approach* ), namely from the effectiveness of classifiers. In other words, as described above, the more well-performing classifiers a gene is included in during Monte Carlo simulations, the higher its weight.

Thus, in the presented study, instead of the common approach (from fundamental observations of transcriptome changes in various conditions to creating a classifier for the purposes of applied medicine), we propose to go in the opposite direction: from effectively working classifiers to understanding pathogenetic processes leading to changes in the transcriptome, which are captured by these classifiers.

To demonstrate the proposed approach, open data of transcriptome profiling of patients with hypertrophic cardiomyopathy (HCM) were selected: GSE36961 and GSE1145.

## METHODS

**Fig. 1**. Study design.

Briefly, at the first stage, we begin with downloading and preprocessing the GSE36961 dataset according to the standard protocol [5] . For training classifiers, we form a data matrix of size $n \times m$ , where $n$ is the number of observations, $m$ is the number of features/genes; the dependent variable is a vector of (0, 1), where 0 means absence of HCM, 1 means presence of HCM. The classification task is set to learn to predict "presence of HCM" based on a feature vector (gene expression levels).

To search for genes involved in the pathogenesis of HCM, we used the Monte - Carlo method to simulate L1-regularized classifiers based on logistic regression. L1-regularization allows thinning of the feature space, leaving only the most significant features (genes) in the classification model. Using this property, we will perform feature selection. Then we trained 3000 models (conducted 3000 simulations), extracting the training sample according to the sampling with replacement scheme. Only observations (rows) were extracted. Genes (features, columns) were not extracted. Each observation (row) was extracted with replacement with equal probability and independently. The test sample was

formed from observations that did not make it into the training sample. As a result, the training and test samples were formed in an approximate ratio of 8:2. Thus, we do not rely on a single model, but simulate many different experiments on various samples obtained by extracting the original sample.

Before launching the algorithm, the regularization coefficient was selected to minimize the decrease in model quality according to the ROC-AUC metric. The coefficient selection and quality assessment were carried out on a labeled training dataset using cross-validation. Thus, we allow overfitting but retain the maximum number of genes, based on the idea that unreliable features will be less frequently included in the model, which will directly affect their weight.

Based on the trained models, we compiled a set of selected genes, which were assigned weights according to the following formula:

$$weight_{gene_j} = \frac{\sum(I_{gene_j \in model_i})}{n} \cdot \frac{\sum(ROCAUC_i \cdot I_{gene_j \in model_i})}{\sum(I_{gene_j \in model_i})} \quad (1)$$

where: $I_{gene_j \in model_i}$ – indicator of the inclusion of the j-th gene in the i-th model, $ROCAUC_i$ – ROC-AUC metric for the i-model, $n$ – number of iterations.

Thus, the weight is defined as the proportion of models in which the gene was included, multiplied by the ROC-AUC quality metric averaged over these models. The ROC-AUC of the model is included in the calculation of the gene weight to distinguish genes selected in the same number of models but differing in classification quality. Subsequently, we will be interested in genes that are most often included in the best classifiers. In this case, the assigned weight will allow us to relevantly order the list of genes for their subsequent processing. Genes that are part of less than 5% of models and have low weight will be excluded from further consideration.

Validation of the results was performed on an independent dataset ( GSE1145 ), which was not used during training or testing. For association assessment (testing the hypothesis of left or right shift), we used the non-parametric Mann-Whitney test [10] , Benjamini - Hochberg - correction for multiple comparisons [11] , as well as the weighted Benjamini - Hochberg correction for multiple comparisons according to the scheme described in [12] .

Statistical tests were conducted using the SciPy module version: 1.7.3. For model training, testing, and data preprocessing, we used the sklearn module version: 0.24.2 [13] .

The algorithm code is available at: https://github.com/GJOsmak/MolBiol2024 .

## RESEARCH RESULTS

In total, the GSE36961 chip contains 37846 transcripts. After data preprocessing, removal of missing values, multimappers and zero readings, 14830 transcripts remained for analysis. Thus, the initial space dimension was 14830 with a sample size of 145 observations. Using these data, we performed 3000 Monte - Carlo simulations as described in the Methods section.

We assume that if a gene is strongly associated with the studied disease, it will be included in most models regardless of the way the sample is split (iteration number). When evaluating algorithm convergence, we decided to call "most significant genes" those that are included in at least half of the models.

As shown in Fig. 2 *a* , the algorithm converges in terms of the number of most significant genes: after ~2000th iteration, the composition of such genes does not change and converges to six genes ( *MYH6* , *CDC42EP4* , *RASD1* , *PRKCD* , *FCN3* , *ZFP36* ). From Fig. 2 *b* , it is evident that after 2000 iterations, the increase in new genes (green line) and the weight increase of the most significant genes (red line) reach a steady state. The rate of weight increase for the most significant genes exceeds the rate of new gene additions. Therefore, it can be assumed that all genes associated with the studied disease were selected within 2000 iterations. All genes selected afterwards are considered noise and are related more to the way the sample is split rather than to the studied disease.

**Fig. 2.** Results of Monte - Carlo simulations for classifier training. *a* - Algorithm convergence by the size of the most significant genes set; red dashes along the x-axis show the moments when this set's composition changes. *b* - Dynamics of growth depending on the algorithm iteration of the number of selected genes (green line); weights of genes included in more than half of the models (red line); iteration at which the set of most significant genes changed (red vertical dashes along the x-axis). *c* – Histogram of ROC-AUC distribution for ML classifiers in 3000 Monte – Carlo simulations. *d* – Histogram of the distribution of calculated gene weights included in at least one model.

As a result, at least 425 genes were included in one of the 3000 models in various combinations. As can be seen from Fig. 2 *c* , most models have a high ROC-AUC score (greater than 0.9). At the same time, the majority of genes (368 out of 425) are included in less than 5% of models (Fig. 2 *d* ). Based on the assumption that a disease-associated gene will be included in most models, we conclude that the usefulness of these 368 genes for classifiers is related more to the way the sample is split than to the disease. For subsequent analysis, the weights of these genes are set to zero. As a

result, the space of tested hypotheses is reduced to 57 genes, which is 260 times smaller than the original space (14830 genes).

As can be seen from Fig. 3, not all of the six "most significant genes" selected above turned out to be statistically significantly associated with the studied disease. The association was not confirmed for the genes *CDC42EP4* , *PRKCD* , *ZFP36.* On the other hand, Fig. 3 *a* shows that along with the genes *MYH6* , *FCN3* and *RASD1,* the gene $_2$ FC, which fell short by 0.06 weight units to be included in the list of "most significant genes". From Fig. 3 *SERPINA3* is also statistically significantly associated and strongly changes its expression by log *b* it can be seen that not all genes that passed the multiple comparison correction (FDR $_{BH}$ ) passed the weighted correction (FDR $_{wBH}$ ). These genes include *INTU, HEG1, SYF2, NKD2, ASPSCR1*.

**Fig. 3**. Testing hypotheses about the association of selected genes on an independent dataset GSE1145. *a* – Volcano plot comparing gene expression, dot size indicates their Weight $_{ML}$ . *b* – Summary statistics table; only significant (by *p* -value) results are shown. *p* -val $_{MW}$ – *p* -value according to the Mann – Whitney test; FDR $_{BH}$ – Benjamini – Hochberg multiple comparison correction; FDR $_{wBH}$ – weighted Benjamini – Hochberg multiple comparison correction; Weight $_{ML}$ – gene weight reflecting its significance for classification models based on Monte – Carlo simulations; log $_2$ FC – logarithm of the ratio of means.

## DISCUSSION OF RESULTS

In this study, we developed and successfully applied an algorithm based on the Monte – Carlo method for generating robust classifiers and using them to prune the feature space (genes). As a result, the analyzed space was reduced by ~260 times from 14830 to 57 genes, which, after subsequent hypothesis testing for associations, were further reduced to 12 genes: *MNS1, FCN3, CHRDL2, MYH6, CAPN1, CD97, S100A9, PROS1, CHN1, SERPINA3, AP3M2, RASD1* , of which, based on the combination of characteristics (calculated weight, log $_2$ FC, adjusted *p* -value), the most noteworthy are *MYH6* , *FCN3* , *RASD1* and *SERPINA3* .

Most of the models during training demonstrated high ROC-AUC metric indicators (mode = 0.96, Fig. 2 *c* ). On the other hand, most genes were included in less than 5% of models (Fig. 2 *d* ). This result is consistent with the consequences of training models in a high-dimensional space, where it is easy to select such a set of features in whose space a particular sample will be well separated; however, this would be an artifact rather than a valuable result [3] .

The gene *MYH6* encodes the alpha isoform of the cardiac myosin heavy chain (α-MHC), which is expressed throughout the myocardium in early stages of heart development. As the human embryo develops, the expression of the *MYH6* gene in the ventricles decreases and is replaced by the expression of *MYH7* [14] . Several studies have shown an association of this gene with HCM [15, 16] .

The product of the gene *FCN3* is a powerful activator of the lectin pathway of complement [17] , associated, according to [18, 19] , with heart failure and ischemic cardiomyopathy [20] .

The monomeric protein RASD1 is expressed in cardiac tissue at a low level [21] . Knockdown of the *RASD1* gene in atrial cardiomyocytes leads to a significant increase in the expression of atrial natriuretic factor [22, 23] , however, no associations of RASD1 with cardiomyopathies have been identified to date.

SERPINA3, also called α-1-antichymotrypsin (AACT, ACT), is one of the serine protease inhibitors, particularly cathepsin G [24] . As an acute phase protein secreted into plasma by liver cells, SERPINA3 plays an important role in the anti-inflammatory response and antiviral response. Elevated levels of SERPINA3 are observed in heart failure and neurological diseases [25] .

Thus, some of the genes discovered using the proposed algorithm are directly related to the disease under study, while others are indirectly related, i.e., the obtained results do not contradict published data. It is also worth noting that the same datasets, GSE36961 and GSE1145, are analyzed in the work [26] , using "standard" approaches, and they arrive at a similar set of genes: *RASD1, CDC42EP4, MYH6* and *FCN3* . Thus, our proposed approach corresponds well with the results of standard approaches, and its advantage lies in the possibility of complete algorithmization and a minimal number of arbitrary decisions. In addition, based on the results of our analysis, another parameter for assessing the "significance" of genes is added – weight. Options for its use are shown in Fig. 3.

## CONCLUSION

In our work, a new algorithm for analyzing transcriptome profiling data is proposed. The results of the algorithm are in good agreement with published data and open up new possibilities for analysis through the generation of a weighted feature space (genes), in contrast to the "standard" situation where all features (genes) are considered as "equal".

## FUNDING

This work was supported by RSF grant No. 23-75-01050.

## ETHICS DECLARATION

The work was carried out without involving people and animals as research objects.

## CONFLICT OF INTERESTS

The authors declare no conflict of interest.

## REFERENCES

1. Akond Z., Alam M., Mollah Md.N.H. (2018) Biomarker identification from RNA-seq data using a robust statistical approach.*Patin S. A.* **http://hdl.handle.net/123456789/1478** (4), 153–163.

2. Tang M., Sun J., Shimizu K., Kadota K. (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data.*No. 4. P. 5* **14.** (1), 360.

3. Barbiero P., Squillero G., Tonda A. (2020) Modeling generalization in machine learning: a methodological and computational study.*arXiv.***2006**Rudneva I.I., Shaida V.G., Medyankina M.V., Shaida O.V. *Assessment of drilling fluid effects on eelgrass*

4. Robinson M.D., McCarthy D.J., Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.*Sedykh V.N., Ignatiev L.A., Semenyuk M.V.* **26**(1), 139–140.

5. 79. *Bioinformatics and computational biology solutions using R and Bioconductor.* doi 10.24412/1728-323X-2021-3-75-79

6. Environmental and public health effects of spent drilling fluid: an updated systematic review // J. Hazard. Mater. Adv . **100120** . https://doi.org/10.1016/j.hazadv.2022.100120 418.

7. Holm S. (1979) A simple sequentially rejective multiple test procedure.*Scandinavian J. Statistics.* **6**(2), 65–70.

8. Gui J., Tosteson T.D., Borsuk M. (2012) Weighted multiple testing procedures for genomic studies.*BioData Mining.* **5**(1), 4.

9. Basu P., Cai T. T., Das K., Sun W (2018) Weighted false discovery rate control in large-scale multiple testing.*J. Am. Stat. Assoc.* **113**(523), 1172–1183.

10. Mann H.B., Whitney D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other.*Ann. Mathemat. Statistics.* **18**(1), 50–60.

11. Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc.: Series B (Methodological).* **57**(1), 289–300.

12. Genovese C.R., Roeder K., Wasserman L. (2006) False discovery control with *p* -value weighting. *Biometrika.* **93** (3), 509–524.

13. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Duchesnay E. (2011) Scikit-learn: machine learning in python. *J. Machine Learning Res.* **12** (Oct), 2825–2830.

14. Anfinson M., Fitts R.H., Lough J.W., James J.M., Simpson P.M., Handler S.S., Mitchell M.E., Tomita-Mitchell A. (2022) Significance of α-myosin heavy chain (MYH6) variants in hypoplastic left heart syndrome and related cardiovascular diseases. *J. Cardiovascular Dev. Dis.* **9** (5), 144.

15. Ntelios D., Meditskou S., Efthimiadis G., Pitsis A., Zegkos T., Parcharidou D., Theotokis P., Alexouda S., Karvounis H., Tzimagiorgis G. (2022) α-Myosin heavy chain (MYH6) in hypertrophic cardiomyopathy: prominent expression in areas with vacuolar degeneration of myocardial cells. *Pathol. Int.* **72** (5), 308–310.

16. Suzuki T., Saito K., Yoshikawa T., Hirono K., Hata Y., Nishida N., Yasuda K., Nagashima M. (2022) A double heterozygous variant in *MYH6* and *MYH7* associated with hypertrophic cardiomyopathy in a Japanese family. *J. Cardiol. Cases.* **25** (4), 213–217.

17. Michalski M., Świerzko A.S., Pągowska-Klimek I., Niemir Z.I., Mazerant K., Domżalska-Popadiuk I., Moll M., Cedzyński M. (2015) Primary ficolin-3 deficiency – is it associated with increased susceptibility to infections? *Immunobiology.* **220** (6), 711–713.

18. Prohászka Z., Munthe-Fog L., Ueland T., Gombos T., Yndestad A., Förhécz Z., Skjoedt MO, Pozsonyi Z., Gustavsen A., Jánoskuti L., Karádi I., Gullestad L., Dahl C.P., Askevold E.T., Füst G., Aukrust P., Mollnes T.E., Garred P. (2013) Association of ficolin-3 with severity and outcome of chronic heart failure. *PLoS One.* **8** (4), e60976.

19. Li D., Lin H., Li L. (2020) Multiple feature selection strategies identified novel cardiac gene expression signature for heart failure. *Front. Physiol.* **11** , 604241.

20. Song H., Chen S., Zhang T., Huang X., Zhang Q., Li C., Chen C., Chen S., Liu D., Wang J., Tu Y., Wu Y., Liu Y. (2022) Integrated strategies of diverse feature selection methods identify aging-based reliable gene signatures for ischemic cardiomyopathy. *Front. Mol. Biosci.* **9** , 805235.

21. Wie J., Kim B.J., Myeong J., Ha K., Jeong S.J., Yang D., Kim E., Jeon J.H., So I. (2015) The roles of Rasd1 small G proteins and leptin in the activation of TRPC4 transient receptor potential channels. *Channels.* **9** (4), 186–195.

22. Kemppainen R.J., Behrend E.N. (1998) Dexamethasone rapidly induces a novel *Ras* superfamily member-related gene in AtT-20 cells. *J. Biol. Chem.* **273** (6), 3129–3131.

23. McGrath M.F., Ogawa T., De Bold A.J. (2012) Ras dexamethasone-induced protein 1 is a modulator of hormone secretion in the volume overloaded heart. *Am. J. Physiol. Heart Circ. Physiol.* **302** (9), H1826–H1837.

24. Baker C., Belbin O., Kalsheker N., Morgan K. (2007) SERPINA3 (aka alpha-1-antichymotrypsin). *Front. Biosci.* **12** (8–12), 2821–2835.

25. de Mezer M., Rogaliński J., Przewoźny S., Chojnicki M., Niepolski L., Sobieska M., Przystańska A. (2023) SERPINA3: stimulator or inhibitor of pathological changes. *Biomedicines.* **11** (1), 156.

26. You H., Dong M. (2023) Prediction of diagnostic gene biomarkers for hypertrophic cardiomyopathy by integrated machine learning. *J. Int. Med. Res.* **51** (11), 03000605231213781.