ACCSCIENCE PUBLISHING

INNOSC Theranostics and Pharmacological Sciences

SHORT COMMUNICATION

Transcriptomic signature of CD4-expressing T-cell abundance developed in healthy peripheral blood predicts strong anti-retroviral therapeutic response in HIV-1: A retrospective and proof-of-concept study

Youdinghuan Chen[1,2,3]*

[1]National Coalition of Independent Scholars (NCIS), Battleboro, Vermont, United States of America

[2]Faculty of Biological and Environmental Informatics, Wilmington University, New Castle, Delaware, United States of America

[3]Department of Science, Mathematics, and Biotechnology, University of California-Berkeley Extension, Berkeley, California, United States of America

Abstract

CD4-expressing T-cells (CD4Ts) play a crucial role in maintaining the normal functioning of the mammalian immune system and overall systemic health. Diseased individuals, such as those infected with the human immunodeficiency virus type I (HIV-1), experience progressive and eventual depletion of CD4T leading to uncurable conditions and ultimate death if left untreated. Although much is known about the role of CD4T-mediated immunity, the understanding of CD4T-related transcriptomic patterns remains incomplete. This proof-of-concept study aims to identify a transcriptome-wide gene signature for CD4T abundance by Least Absolute Shrinkage and Selection Operator (LASSO) regression modeling in 340 healthy peripheral blood samples. The optimized LASSO model demonstrated computational robustness (tenfold average Pearson's r = 0.89) and biological relevance evidenced by four significant Gene Ontology terms (all odds ratio [OR] ≥ 4.5 and false discovery rate ≤0.05). Subsequently, in an independent cohort with 24 HIV-1-infected men who received anti-retroviral therapies, there is a significant, positive association between the gene signature and a strong anti-retroviral response before (OR = 13.6, $P < 0.05$) and after adjusting for subject age, sex, and race (OR = 14.4, $P < 0.05$). Taken together, the gene expression pattern associated with CD4T abundance is predictable, generalizable, and biologically relevant, shedding new light on the importance of CD4T abundance.

Keywords: CD4T; Gene signature; HIV-1; Immune cells; Transcriptomics

***Corresponding author:**
Youdinghuan Chen
(y.david.chen@ncis.org)

Citation: Chen Y. Transcriptomic signature of CD4-expressing T-cell abundance developed in healthy peripheral blood predicts strong anti-retroviral therapeutic response in HIV-1: A retrospective and proof-of-concept study. *INNOSC Theranostics and Pharmacological Sciences.* 2024;7(3):2761. doi: 10.36922/itps.2761

Received: January 17, 2024

Accepted: May 16, 2024

Published Online: July 12, 2024

Copyright: © 2024 This is an Open-Access article distributed under the terms of the Creative Commons AttributionNoncommercial License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations

1. Introduction

CD4-expressing T lymphocytes (CD4Ts) are crucial players in the human immune system. The CD4T plays a role in the modulation of antiviral immune responses primarily by enhancing the function of other immune cells. During the initial stages

of a viral infection, CD4 helper T-cells activate B-cells and cytotoxic lymphocytes to enable the proper immune response against the invading pathogen.[1]

In the case of human immunodeficiency virus (HIV) infection, CD4Ts are specifically targeted, exploited, and destroyed, resulting in compromised immunity.[1,2] The progressive depletion of CD4T, as seen in HIV-infected patients without anti-retroviral treatment, is responsible for the development of acquired immunodeficiency syndrome, which renders an individual vulnerable to even the most commonplace opportunistic pathogens.[1,2] Among HIV patients receiving anti-retroviral treatment, CD4T abundance is associated with favorable clinical outcomes and serves as a prognosis indicator.[3]

The most recent meta-analysis of transcriptomes validated the hypothesis that HIV infection induces characteristic changes in CD4T gene expression and biological pathways.[4] Nevertheless, an in-depth understanding of CD4T biology in both the healthy and diseased states remains limited. Thus, this study aims to explore the possibility of building a transcriptome-wide gene signature using clinically measured CD4T abundance as the outcome in blood samples of non-diseased human subjects. Subsequently, this study investigates the potential prognostic utility of the gene signature in a cohort of HIV-1-positive subjects receiving anti-retroviral therapies to offer insights into the role of CD4T abundance.

2. Materials and methods

2.1. Study populations and datasets

All datasets used in this study are publicly available and obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo). The "discovery population" (accession GSE58137) consists of 340 subjects with both transcriptome-wide gene expression data (measured by the Illumina Human HT 12 v3.0-4.0 array) and the putative CD4T proportions calculated from the matched whole-genome methylation profiles.[5] The "application population" (accession GSE19087) consists of 24 HIV-1 positive men with both gene expression (Illumina Human WG 6 v3.0 array) and CD4T counter before and after a 48-week anti-retroviral therapeutic regimen.[3] Covariates including age and race were available in both populations. Subject demographics are summarized in Table S1.

Some genes measured on the Illumina array had multiple transcript variants. In addition, the expression platforms were of two different versions. To ensure comparability between the datasets, each population's expression values were aggregated by unweighted mean across all available transcript variants of a common gene so that there would be only one unique gene symbol. Next, each population's gene expression matrix was restricted to a common set of 12,549 genes. To remove platform-related artifacts or experimental noise, each expression matrix was mean-centered and standard deviation-scaled and then constrained within values ±5.0 as previously described.[6-8]

2.2. Gene signature discovery from the transcriptomes of healthy peripheral blood

The CD4T abundance gene signature was identified using Least Absolute Shrinkage and Selection Operator (LASSO; *glmnet* R package v.4.1.8), an established approach with the ability to reduce model coefficients to zero, thereby nullifying the effect estimate of a given feature unless it is very strong.[9,10] LASSO was built on the shared 12,549-gene set against the min-max scaled CD4T proportions in the discovery population, as follows:

$$MinMax(CD4T) \sim gene_{k} + R \quad \text{(I)}$$

where R is the LASSO penalty term defined by setting the *glmnet* hyperparameter alpha to 1.0. *MinMax* is a scaling function that transforms the input vector into a distribution between 0 and 1:

$$MinMax(x_{k}) = (x_{k} - min(\mathbf{x}))/(max(\mathbf{x})-min(\mathbf{x})) \quad \text{(II)}$$

This procedure initially yielded 334 (2.7%) gene features with non-zero regression coefficients. Given some coefficients had low magnitudes, the LASSO-selected genes were further filtered by a coefficient threshold of 1.25e-3, yielding 207 (114 positive and 93 negative) gene features, hereafter referred to as the "CD4T gene signature" (Table S2).

2.3. *In silico* validation of the gene-signature final model with *K*-fold method

The identified gene signature was validated with a *K*-fold cross-validation experiment. Specifically, the entire dataset was divided into K=10 partitions (i.e., folds, each with n = 34). The final, optimized version of the LASSO model was built on K-1=9 training partitions and evaluated on the remaining test partition. For each round of cross-validation, several metrics for model evaluations were calculated on the test partition: Root mean-squared error, Pearson's correlation coefficient (r), and R-squared. The entire procedure was repeated K times so that every fold participated in at least one round of training and testing. Each evaluation metric was averaged across the K trials to yield a final performance indicator (Table S1).

2.4. Biological interpretation of the CD4T gene signature

Separate *Gene Ontology: Biological Processes no-redundant* analyses (WebGestalt implementation, www.webgestalt.org)[11] were performed on the gene sets with positive and

negative LASSO coefficients against the 12,549 genes used as the input for gene-signature discovery. Raw *P*-values were adjusted by the Benjamini–Hochberg false discovery rate (FDR) method.

2.5. Gene signature-based stratification of an independent cohort of HIV-1-positive men

The CD4T gene signature was subsequently applied to the application population consisting of 24 peripheral blood samples from HIV-1-infected men receiving anti-retroviral therapies. All study participants had the absolute CD4T cell count before and after treatment. The percent change is defined as:

$$\text{Percent change} = 100 \times \frac{\text{Change in CD4T}}{\text{Baseline CD4T}} \quad \text{(III)}$$

Unsupervised hierarchical clustering with Euclidean and *Ward D* hyperparameters (*pheatmap* R package v.1.0.12) followed by a dendrogram-tree split at the first node was used to stratify the application population into two groups for downstream statistical analyses.

2.6. Statistical analysis

Unless otherwise specified, the computational environment used was R 4.3.1 (https://www.r-project.org) with data analysis and visualization packages *base* (4.3.1), *ggplot2* (v.3.4.3), and *matrixStats* (v.1.0.0). All statistical tests used were two-sided. The univariate (unadjusted) association between two binary variables was determined by a Fisher's exact test with odds ratio (OR) and 95% confidence interval (CI) estimates. To address potential confounding, the multivariate association was determined by multivariate logistic regression (a generalized linear model with a family "Binomial") with adjusted ORs estimated by exponentiating the model coefficients. The mean difference between any two groups was determined by Welch's *t*-test. A generalized linear model with family "Gaussian" was the generalization of the *t*-test to control for potential confounding variables. All R code is deposited to GitHub (https://github.com/ydavidchen/cd4t_pilot_signature).

3. Results

3.1. Identification and characterization of a CD4T abundance signature in the transcriptomes of healthy peripheral blood samples

A transcriptomic gene signature was identified by supervised modeling of gene expression against CD4T proportions using LASSO regression, a conservative statistical approach for selecting the most important features from high-dimensional data space.[9,10] LASSO is well-known for its ability and tendency to shrink the model coefficients to zero, thereby revealing input features (i.e., genes) most strongly correlated with the outcome (i.e., CD4T abundance). A tenfold cross-validation experiment demonstrated the robustness of the final model (tenfold average Pearson's $r = 0.89$ and $r^2 = 0.79$; Table S1). The initial CD4T gene signature consisted of 334 (2.7%) genes with non-zero LASSO coefficients (Figure 1A). The final version of the signature with 207 genes (1.6%) was obtained by coefficient thresholding (Figure 1A and Table S2).

To explore the biological relevance of the identified gene signature, *Gene Ontology: Biological Processes* analysis was performed on the gene features in the positive and negative directions, separately. The gene features positively associated with CD4T abundance strongly enriched for cellular adhesion (OR = 4.7, FDR = 0.01; Table 1). Notably, the members comprising this ontology term included high-profile immune genes involved in HIV-1 pathogenesis: CD28 and CTLA-4. The gene set negatively associated with CD4T abundance strongly enriched for metabolic processes of macromolecules (all OR ≥ 4.5 and all FDR = 0.05; Table 1). The genes encoding CD8 subunits, *CD8A* and *CD8B*, showed strong, negative association with CD4 abundance and were selected by the LASSO procedure (Table S3).

Hierarchical clustering of the identified gene signature segregated the discovery population into two major clusters: *Cluster 1* and *Cluster 2* (Figure 1B). The distribution of subject demographics, including race/ethnicity, sex, and age, appeared balanced across the gene-signature clusters (Figure 1B**,** horizontal tracking bars).

3.2. Application of CD4T gene signature to an HIV-1-positive cohort for biomedical knowledge discovery

The next objective was to assess the clinical relevance of the CD4T gene signature in human disease. Given the well-known role of CD4T in HIV-1 infection and recovery on anti-retroviral treatment,[3] the CD4T gene signature was evaluated in this disease context. Dataset GSE19087 has 24 HIV-1 positive men treated with an anti-retroviral regimen for approximately 1 year.[3] Hierarchical clustering of the CD4T gene signature stratified the HIV-1 positive, anti-retroviral treated men into two major groups (Figure 2). The cluster structure, indicated by the pattern of dendrogram branching, showed striking similarity to that of the discovery population. On stratification of cluster membership, the subject demographics present no significant differences between the two clusters (Table 2). However, the HIV-1-positive men in *Cluster 1* had an average of 122.7% CD4 increase at the end of the anti-retroviral therapy treatment, compared to

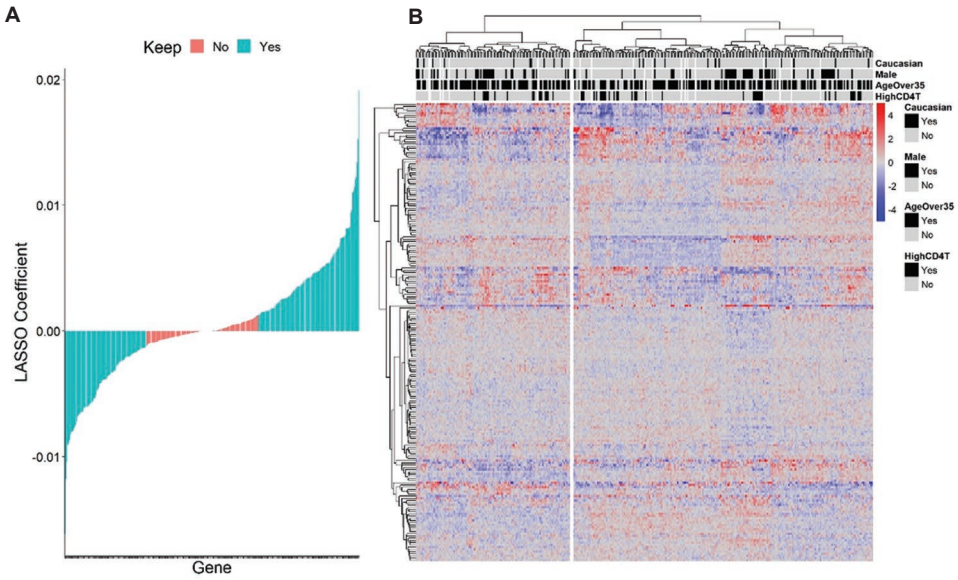


Figure 1. Discovery of the gene signature. (A) Distribution of 334 non-zero LASSO coefficients of the detected genes, with 207 of which meeting a 0.0015 cutoff included for downstream analyses. (B) Heat map showing the normalized expression of the 207 genes as rows and the 340 samples as columns. Horizontal tracking bars indicate clinical covariates. The dendrogram denotes the result of unsupervised hierarchical clustering.
Note: "HighCD4T" indicates whether a sample has an estimated CD4T proportion over 25%.

Table 1. ***Gene Ontology: Biological Processes*** **for the gene signature members, either positively or negatively associated with CD4T abundance in the discovery population**

Association with CD4T	ID	Process name	Number of genes observed	Number of genes expected	OR	*P*-value	FDR
Positive	GO: 0045785	Positive regulation of cell adhesion	11	2.3	4.7	1.7e-5	0.01
Negative	GO: 1901615	Organic hydroxy compound metabolic process	9	2.0	4.5	1.5e-4	0.05
Negative	GO: 0046486	Glycerolipid metabolic process	9	1.9	4.8	8.6e-5	0.05
Negative	GO: 0008202	Steroid metabolic process	7	1.2	5.8	1.7e-4	0.05

Abbreviations: OR: Odds ratio; FDR: False discovery rate.

Cluster 2 which had only 32.7%. On average, the CD4T percent increase was 90% higher in *Cluster 1* than in *Cluster 2 (*95% CI = 34.9 – 145.1%, *P* = 0.003, Welch's *t*-test; Table 2). This difference remained consistent and statistically significant after adjusting for subject age and race (estimated mean difference 81.9%, 95% CI = 4.6 – 159.2%, *P* = 0.05, multiple regression; Table 3A).

In the application population, an HIV-1 positive man was defined as having a "strong antiviral response (SAR)" if he showed a minimum of 50% increase in CD4T cell count by the end of the treatment regimen. All except one SAR subject were assigned to *Cluster 1*, the majority cluster (Figure 2 – horizontal tracking bars, and Table 2). On average, the SARs were 13.6 times more likely to adopt a *Cluster 1* gene signature pattern (95% CI = 1.2 – 751.3, *P* = 0.027, Fisher's exact test; Table 2). To rule out the possibility that known confounders drove the observed association, a multivariate logistic regression model was built. Adjusting for age and race, the association between the strong response and *Cluster 1* assignment remained significantly positive (adjusted OR = 14.4, 95% CI = 1.7 – 337.4, *P* = 0.03; Table 3B).

4. Discussion

This study provides evidence that transcriptome-wide gene expression captures biological and clinical insights into an infectious disease based on healthy CD4T abundance

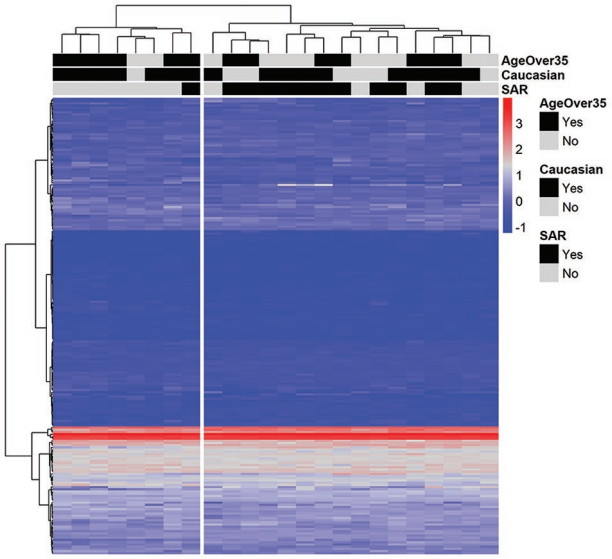


Figure 2. Application of the gene signature to an independent population of HIV1-positive, anti-retroviral therapy-treated men. Each row and column represents a gene and a sample, respectively. Horizontal tracking bars at the top indicate clinical covariates, which were used in univariate and/or multivariate analyses.
Note: "SAR" indicates a strong anti-retroviral response.

profiles. Using a statistically robust and conservative approach, LASSO, a panel of CD4T genes capable of stratifying high- versus low-abundance samples, was identified. The Gene Ontology analysis revealed that regulation of cell adhesion was positively associated with CD4T abundance. This result was unsurprising. During an inflammatory response, immune cells must be able to migrate and localize to the target tissue. *CD28* was among the genes accounting for this Gene Ontology term and was also top-ranked among all genes with positive LASSO coefficients. CD28 is a cell-surface glycoprotein essential for T-cell survival and proliferation. The implication of CD28 in HIV-1 pathogenesis has long been recognized. In HIV-1 patients, CD4 and CD28 coexpression is a predictor of progression to acquired immunodeficiency syndrome (AIDS).[12] Mechanistically, CD28 expressed on the T-cell surface can be targeted by two HIV-1 accessory proteins *Nef* and *Vpu* for degradation.[13] Meanwhile, the CD4T gene signature was depleted for metabolic processes of bio-macromolecules including glycerol-lipids and steroids. Unsurprisingly, both subunits of the CD8 molecule (*CD8A* and *CD8B*) were downregulated since the machine-learning model was optimized against characteristics associated with CD4 abundance. In a healthy tissue context, CD4T cells undergo metabolic programming and adaptations during an inflammatory response. Moreover, T-cell activation requires extensive metabolic reprogramming and elevated energy expenditure involving

Table 2. Stratification of application population by the gene-signature clusters

	Cluster 1	Cluster 2	*P*-value
n (%)	16 (66.7)	8 (33.3)	-
Age (mean±SD)	34.25±9.59	41.88±10.09	0.10
Race (%)			0.35
Caucasian	10 (62.5)	7 (87.5)	
Other	6 (37.5)	1 (12.5)	
CD4% change (mean±SD)	122.66±94.21	32.67±34.31	0.003**
SAR (%)			0.027*
Yes	11 (68.8)	1 (12.5)	
No	5 (31.2)	7 (87.5)	

Notes: *P* values were determined by two-sided Fisher's exact and Welch's *t*-test for discrete and continuous variables, respectively. *$P<0.05$, **$P<0.01$.
Abbreviation: SAR: Strong anti-viral response.

Table 3. Multivariate-adjusted analyses

(A)	Adjusted OR (95% CI)	*P*-value
SAR		
No	1.0 (reference)	-
Yes	14.4 (1.6 – 337.4)	0.03*
Age in years	0.94 (0.82 – 1.1)	0.32
Caucasian race		
No	1.0 (reference)	-
Yes	0.33 (0.01 – 4.3)	0.43
(B)	**Mean percent increase in CD4T level (95% CI)**	***P*-value**
Gene signature cluster		
#2	0.0 (reference)	-
#1	81.9 (4.6 – 159.2)	0.05*
Age in years	−0.51 (−4.1 – 3.1)	0.79
Caucasian race		
No	0.0 (reference)	-
Yes	−16.71 (−93.8 – 60.4)	0.68

Notes: Part (A) represents the association between strong antiviral response (SAR) with Gene Signature Cluster 1 estimated by multivariate logistic regression. Adjusted ORs were determined by exponentiating the model coefficients. Part (B) represents the mean difference between Gene Signature Clusters 1 and 2 estimated by multivariate linear regression. *$P<0.05$.
Abbreviations: OR: Odds ratio; CI: Confidence interval.

the aforementioned bio-macromolecules.[14] This finding, which appears contradictory, might be explained by a negative feedback loop due to overabundant CD4T.

When applied to an HIV-1-positive patient cohort receiving anti-retroviral therapies, the identified CD4T gene signature stratified the study population into clusters

significantly related to clinical response. The near-perfect stratification of diseased subjects here is consistent with the prior knowledge that CD4T abundance is essential to a favorable prognosis of anti-retroviral treatment response in HIV-1 patients.[1,2] In addition, formal statistical testing revealed a significant association between the gene signature and anti-retroviral response. This observed effect remained consistent in direction and significant statistically after accounting for clinical covariates. Taken together, these results support the generalizability of CD4T abundance-based gene signature, identified in the healthy context, to the diseased context, and reinforce the centrality of CD4T-based immune modulation during HIV-1 pathogenesis.

Several limitations of this study should be acknowledged. First, despite having high-quality gene-expression data, the application population used for biological and clinical inference is small. In particular, the LASSO model performance might improve if model training was conducted in a sex-stratified manner, a known risk-modifying factor in HIV/AIDS.[15] Due to the limited training sample size, model development on sex-stratified data was unlikely empowered to achieve robust performance. Second, there were variations in the study designs, and many clinical variables of the study populations used were not consistently collected or reported. Importantly, the gene-signature identification process leverages only gene expression without clinical covariates such as subject age and race. Finally, a broader limitation is the availability of datasets and samples with high consistency (i.e., low heterogeneity). Future confirmatory and follow-up studies with comprehensive and robust designs will be crucial to the understanding of the molecular mechanisms of diseases related to CD4T abundance.

Acknowledgments

The author would like to thank the anonymous peer-reviewers and the journal editorial team for their input.

Funding

None.

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

As described in the Materials and Methods section, the datasets and code used in this study can be publicly accessed through Gene Expression Omnibus and GitHub, respectively.

References

1. Swain SL, McKinstry KK, Strutt TM. Expanding roles for CD4+ T cells in immunity to viruses. *Nat Rev Immunol*. 2012;12(2):136-148.

   doi: 10.1038/nri3152

2. Vidya Vijayan KK, Karthigeyan KP, Tripathi SP, Hanna LE. Pathophysiology of CD4+ T-cell depletion in HIV-1 and HIV-2 Infections. *Front Immunol*. 2017;8:580.

   doi: 10.3389/fimmu.2017.00580

3. Woelk CH, Beliakova-Bethell N, Goicoechea M, *et al*. Gene expression before HAART initiation predicts HIV-infected individuals at risk of poor CD4+ T-cell recovery. *AIDS*. 2010;24(2):217-222.

   doi: 10.1097/QAD.0b013e328334f1f0

4. Coelho AVC, Gratton R, de Melo JPB, *et al*. HIV-1 Infection transcriptomics: Meta-analysis of CD4+ T cells gene expression profiles. *Viruses*. 2021;13(2):244.

   doi: 10.3390/v13020244

5. Zannas AS, Jia M, Hafner K, *et al*. Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk. *Proc Natl Acad Sci U S A*. 2019;116(23):11370-11379.

   doi: 10.1073/pnas.1816847116

6. Chen Y. A cancer proliferation gene signature supervised by Ki-67 strata specific to luminal A, estrogen receptor-positive, and HER2-negative ductal carcinomas. *Med Res Arch*. 2022;10(10).

   doi: 10.18103/mra.v10i10.3160

7. Chen Y. Transcriptomic profiling of subcutaneous adipose tissue in relation to bariatric surgery: A retrospective, pooled re-analysis. *J Obes Metab Syndr*. 2023;32(1):98-102.

   doi: 10.7570/jomes22065

8. Chen Y. Pooled microarray expression analysis of failing left ventricles reveals extensive cellular-level dysregulation independent of age and sex. *J Mol Cell Cardiol Plus*. 2024;7:100060.

   doi: 10.1016/j.jmccpl.2023.100060

9. Tibshirani R. Regression selection and shrinkage via the lasso. *J R Stat Soc B*. 1996;58(1):267-288.

   doi: 10.2307/2346178

10. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.

11. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199-W205.

doi: 10.1093/nar/gkz401

12. Choremi-Papadopoulou H, Panagiotou N, Samouilidou E, *et al*. CD28 costimulation and CD28 expression in T lymphocyte subsets in HIV-1 infection with and without progression to AIDS. *Clin Exp Immunol*. 2000;119(3):499-506.

doi: 10.1046/j.1365-2249.2000.01153.x

13. Pawlak EN, Dirk BS, Jacob RA, Johnson AL, Dikeakos JD. The HIV-1 accessory proteins Nef and Vpu downregulate total and cell surface CD28 in CD4+ T cells. *Retrovirology*. 2018;15(1):6.

doi: 10.1186/s12977-018-0388-3

14. Dumitru C, Kabat AM, Maloy KJ. Metabolic adaptations of CD4+ T cells in inflammatory disease. *Front Immunol*. 2018;9:540.

doi: 10.3389/fimmu.2018.00540

15. Scully EP. Sex differences in HIV infection. *Curr HIV/AIDS Rep.* 2018;15(2):136-146.

doi: 10.1007/s11904-018-0383-2