ORIGINAL ARTICLE

DOI: 10.4274/nkmj.galenos.2026.65807
Nam Kem Med J 2026;14(2):160-171

Cross Platform Blood Transcriptomics Identifies a Two-*gene* Classifier for Coronary Artery Disease Detection

Koroner Arter Hastalığının Saptanmasında Platformlar Arası Kan Transkriptomi Analiziyle İki Genli Bir Sınıflandırıcının Belirlenmesi

Bilge Eren YAMASAN[1], Selçuk KORKMAZ[2]

[1]Trakya University Faculty of Medicine, Department of Biophysics, Edirne, Türkiye
[2]Trakya University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Edirne, Türkiye

ABSTRACT

Aim: Coronary artery disease (CAD) remains a major global health burden, and currently available blood biomarkers lack sensitivity for early detection. This study aimed to identify and validate circulating mRNA and long non-coding RNA (lncRNA) biomarkers of CAD and to develop a predictive transcriptomic model.

Materials and Methods: We analyzed plasma RNA-seq data from stable CAD patients and healthy controls (*GSE208194*) and validated findings in an independent peripheral blood microarray cohort (*GSE113079*). Differential expression was assessed using a threshold of $|\log^2$ fold-change$| \geq 1$ and false-discovery rate <0.05. Enrichment analysis of gene ontology and Kyoto Encyclopedia of Genes and Genomes pathways was performed. A predictive model was constructed using logistic regression model-penalized logistic regression, with hyperparameters tuned within a nested cross-validation framework, and evaluated in the independent validation cohort.

Results: A total of 182 transcripts (177 mRNAs, 5 lncRNAs) were differentially expressed, with 91% down-regulated in CAD. Enrichment analysis revealed coordinated dysregulation of ribosomal biogenesis, cytoplasmic translation, mitochondrial oxidative phosphorylation, and p53/NF-κB inflammatory pathways. Cross-platform validation confirmed 85 transcripts, indicating a robust expression signature. The predictive model selected two genes, *NEUROD2* and *RPS27*, achieving an external area under the curve of 0.820 in the validation cohort.

Conclusion: Blood transcriptomic profiling identifies a reproducible CAD-associated expression signature and supports a concise two-gene classifier reflecting inflammatory and metabolic stress-related pathways. These findings provide a framework for the further development of blood-based transcriptomic assays and warrant validation in larger, diverse populations to define clinical utility in cardiovascular risk assessment.

Keywords: Coronary artery disease, transcriptome, biomarkers, RNA, logistic models

ÖZ

Amaç: Koroner arter hastalığı (KAH) küresel ölçekte önemli bir sağlık yükü olmaya devam etmektedir ve mevcut kan biyobelirteçleri erken tanı için yeterli duyarlılığa sahip değildir. Bu çalışmanın amacı, KAH ile ilişkili dolaşımdaki mRNA ve uzun kodlamayan RNA (lncRNA) biyobelirteçlerini tanımlamak ve doğrulamak, ayrıca öngörücü bir transkriptomik model geliştirmektir.

Gereç ve Yöntem: Stabil KAH hastaları ve sağlıklı kontrollerden elde edilen plazma RNA-sekanslama verileri (*GSE208194*) analiz edilmiş ve bulgular bağımsız bir periferik kan mikrodizi kohortunda (*GSE113079*) doğrulanmıştır. Diferansiyel gen ekspresyonu $|\ |\log^2$ kat değişimi$| \geq 1$ ve yanlış keşif oranı <0,05 eşikleri kullanılarak değerlendirilmiştir. Gen ontolojisi ve Kyoto Genler ve Genomlar Ansiklopedisi yolakları için zenginleştirme analizleri yapılmıştır. Öngörücü model, lojistik regresyon modeli ile cezalandırılmış lojistik regresyon kullanılarak oluşturulmuş, hiperparametreler iç içe çapraz doğrulama çerçevesinde ayarlanmış ve bağımsız doğrulama kohortunda değerlendirilmiştir.

Address for Correspondence: Selçuk KORKMAZ MD, Trakya University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Edirne, Türkiye
E-mail: selcukorkmaz@gmail.com **ORCID ID:** orcid.org/0000-0002-6525-2503
Received: 08.10.2025 **Accepted:** 01.02.2026 **Publication Date:** 16.06.2026

Cite this article as: Yamasan BE, Korkmaz S. Cross platform blood transcriptomics identifies a two-gene classifier for coronary artery disease detection. Nam Kem Med J. 2026;14(2):160-171

Bulgular: Toplam 182 transkript (177 mRNA, 5 lncRNA) diferansiyel olarak eksprese bulunmuş ve bunların %91'i KAH'de aşağı regüle edilmiştir. Zenginleştirme analizleri ribozomal biyogenez, sitoplazmik translasyon, mitokondriyal oksidatif fosforilasyon ve p53/NF-κB enflamatuvar yolaklarında eşgüdümlü düzensizlikleri ortaya koymuştur. Platformlar arası doğrulama 85 transkripti teyit ederek sağlam bir ekspresyon imzasına işaret etmiştir. Öngörücü model iki geni (*NEUROD2* ve *RPS27*) seçmiş ve doğrulama kohortunda 0,820'lik harici eğri altında kalan alan değerine ulaşmıştır.

Sonuç: Kan transkriptomik profillemesi, KAH ile ilişkili, tekrarlanabilir bir gen ekspresyon imzası ortaya koymakta ve enflamasyon ile metabolik stres süreçlerini yansıtan iki genli, yalın bir sınıflandırıcının kullanılabileceğini göstermektedir. Bu bulgular, kan temelli transkriptomik biyobelirteçlerin geliştirilmesine yönelik önemli bir temel sağlamaktadır. Ancak klinik uygulamadaki değerinin ortaya konabilmesi için daha geniş ve farklı özelliklere sahip hasta popülasyonlarında doğrulanması gerekmektedir.

Anahtar Kelimeler: Koroner arter hastalığı, transkriptom, biyobelirteçler, RNA, lojistik modeller

INTRODUCTION

Coronary artery disease (CAD) remains a major global health problem, responsible for high morbidity and mortality despite preventive and therapeutic advances. Cardiovascular diseases cause over 20 million deaths annually[1] with ischemic heart disease alone accounting for about 9 million[2]. Despite decades of advances in prevention and therapy, CAD continues to be a leading cause of morbidity and mortality globally[3]. CAD arises from atherosclerosis, a chronic inflammatory process triggered by lipid deposition, endothelial dysfunction, and immune activation[4]. These mechanisms, including vascular inflammation and thrombosis, drive its clinical manifestations. However, many individuals harbor subclinical CAD undetected until acute coronary events ocur[5].

Current blood-based biomarkers mainly reflect risk factors or myocardial injury rather than subclinical atherosclerosis. Low-density lipoprotein cholesterol and other lipids miss many at-risk patients[4], while C-reactive protein is nonspecific and troponins rise only after myocardial damage[6]. Imaging techniques like coronary computed tomography (CT) or angiography offer definitive diagnosis but are costly or invasive[7]. Hence, new non-invasive biomarkers reflecting early pathobiological changes are needed.

Transcriptomic profiling enables genome-wide assessment of mRNA and non-coding RNA expression in blood, capturing systemic molecular alterations[8]. Studies have shown distinct blood gene-expression signatures in CAD. The 23-gene Corus CAD score accurately predicted obstructive CAD and adverse outcomes in validation cohorts[9-11]. Long non-coding RNAs (lncRNAs) also regulate lipid metabolism and inflammation and are stable in circulation[4,5,12]. For example, *ANRIL* at the chromosome 9p21 locus links genetic risk to CAD[13], while lncRNAs such as *H19*, *MIAT*, and *MALAT1* are elevated in acute myocardial infarction[14,15].

Nonetheless, most studies lack replication, cross-platform validation, or inclusion of non-coding RNAs. Many derive from single cohorts with limited generalizability[7,8]. The Corus score included only mRNAs and excluded diabetics[9,11], while newer efforts only recently began profiling lncRNAs systematically[16]. Moreover, predictive modeling translating transcriptomic data into clinical tools remains rare, with few classifiers validated beyond the Corus model[9,11,17].

This study aims to identify and validate blood-based mRNA and lncRNA biomarkers for CAD to build a robust diagnostic classifier. By integrating both RNA types and validating across independent populations, it seeks a comprehensive transcriptomic signature to enhance early, noninvasive CAD detection and risk assessment.

MATERIALS AND METHODS

Datasets

This study was a retrospective bioinformatics analysis of two publicly available GEO transcriptomic datasets on CAD. The discovery dataset, *GSE208194*, comprised RNA-seq data from circulating cell-free RNA (cfRNA) in plasma from 59 patients with stable CAD (4 months post-acute event) and 30 healthy controls, sequenced on an Illumina NovaSeq 6000. The validation dataset, *GSE113079*, was a microarray-based transcriptome of PBMCs from 93 CAD patients and 48 controls (Agilent Human lncRNA+mRNA array v4.0), covering genome-wide mRNA and lncRNA expression. *GSE208194* reflects circulating RNA released from tissues, whereas *GSE113079* captures immune-cell transcript levels, providing an independent cohort and platform for validation.

Data Acquisition and Preprocessing

All data were obtained from GEO using the GEOquery package (v2.66.0) in R.[18] for the discovery RNA-seq dataset (*GSE208194*), we downloaded the processed gene-level expression matrix provided by the data submitters, reported as transcripts per million (TPM). While raw sequencing reads for this dataset are available via the Sequence Read Archive (SRA), a gene-level raw count matrix is not included in the GEO Series record. As this study represents a secondary analysis of publicly available data, downstream analyses were therefore conducted using the supplied TPM-based expression matrix rather than reprocessing raw FASTQ files.

Gene identifiers were mapped from Ensembl IDs to official gene symbols using GENCODE v33 human gene annotation[19]. Analyses were restricted to protein-coding genes and annotated lncRNAs. To ensure robust detection and reduce noise from low-abundance transcripts, features with TPM ≥1 in at least 90% of samples were retained. Remaining TPM values were log-transformed [$\log^2$ (TPM + 1)] to stabilize variance. Quality control was performed by inspection of sample-level expression distributions and principal component analysis (PCA).

For the validation dataset *GSE113079* (Agilent one-color microarray), we downloaded the processed expression matrix provided by the submitters, consisting of background-corrected, $\log^2$-transformed probe intensities. As described in the original publication[20], the arrays were quantile-normalized in GeneSpring GX prior to submission. Probe identifiers were mapped to gene symbols using the platform annotation file (*GPL20115*). Analyses were restricted to genes represented in both datasets, enabling direct cross-platform validation. All microarray samples (93 CAD, 48 control) passed quality control in the original study and were retained for downstream analyses.

Differential Expression Analysis

In the discovery plasma RNA-seq dataset (*GSE208194*), differential expression between CAD and control groups was analyzed using the limma package (v3.62.2)[21]. The analysis was performed on $\log^2$ (TPM + 1)–transformed gene-level expression values, with disease status (CAD vs. control) specified as the primary model term. Because all samples were generated within a single sequencing run and no additional technical covariates were reported in the GEO record, no further adjustment variables were included in the linear model.

For each gene, limma's linear modeling framework with empirical Bayes moderation was applied to obtain moderated log-fold changes and associated statistics, improving variance estimation and statistical stability. Genes were considered differentially expressed if they met both an absolute $\log^2$ fold-change threshold of ≥1.0 and a Benjamini–Hochberg false-discovery rate (FDR) <0.05, ensuring selection of transcripts with both statistical support and biologically meaningful effect sizes. The resulting differentially expressed genes (DEG) were retained as candidate circulating RNA biomarkers for CAD. Volcano plots were used to visualize the overall distribution of effect sizes and significance levels and to verify that the selected thresholds captured the most prominent mRNA and lncRNA signals.

Validation in Independent Cohort

To validate the RNA-seq findings, we analyzed the *GSE113079* PBMC microarray dataset as an independent cohort. A similar differential expression analysis was performed using limma on the microarray gene expression matrix. For each gene, a linear model was fitted with group (CAD vs. control) as the contrast of interest, followed by empirical Bayes moderation to obtain moderated statistics. The larger sample size and normalization of the original dataset helped minimize potential batch effects, which were further checked using PCA plots showing no evident batch structure. Differential expression was defined by an FDR <0.05, consistent with the RNA-seq criteria. Genes were considered validated if they were significantly differentially expressed in the microarray cohort (FDR <0.05) with a concordant direction of change between datasets. This cross-platform validation ensured that identified mRNA and lncRNA biomarkers were reproducible across independent cohorts and technologies. Genes not represented on the microarray (e.g., certain lncRNAs) were excluded, and the validated gene set was used for subsequent analyses. In addition to statistical significance and concordant direction of effect, cross-platform effect-size concordance was quantified by calculating Pearson correlations of $\log^2$ fold-change estimates between the discovery RNA-seq and validation microarray datasets. Correlations were computed separately for (i) the validated gene set and (ii) the full set of discovery DEG that were represented on the microarray platform.

Functional Enrichment Analysis

To interpret the biological significance of the DEG, we performed gene ontology (GO) and pathway enrichment analyses using the clusterProfiler package (v4.14.6)[22]. Significantly up- and down-regulated genes were tested for enrichment in GO biological process (BP), molecular function, cellular component, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway categories. Enrichment was assessed using a one-tailed hypergeometric test, with the background defined as all genes expressed above the filtering threshold in the RNA-seq dataset to control for detection bias. Multiple testing correction was applied using the Benjamini-Hochberg procedure, and terms with FDR <0.05 were considered significant. Results were summarized with enrichment fold changes (gene ratios) and adjusted p-values. Visualization of the top enriched GO terms and KEGG pathways was done using clusterProfiler (v4.16.0)[23]. Since lncRNAs are not directly annotated to GO/KEGG terms, enrichment primarily reflected the differentially expressed mRNAs. The results were interpreted in the context of CAD-related processes, including ribosomal biogenesis, oxidative phosphorylation, RNA processing, and DNA damage response.

Predictive Modeling

Using the discovery plasma RNA-seq dataset, we built a predictive model to test whether the identified transcriptomic biomarkers could distinguish CAD patients from controls.

An L1-penalized logistic regression model (LASSO) was applied, suitable for high-dimensional data and automatic variable selection by shrinking some coefficients to zero. To prevent feature-selection leakage, differential expression analysis was not performed globally prior to modeling. Instead, DEG identification was repeated independently within each training fold during cross-validation, and only fold-specific DEGs were used as candidate predictors. Expression data for candidate biomarkers were standardized to zero mean and unit variance within each training fold to ensure strict separation between training and assessment data. The binary outcome was CAD status. All modeling steps were implemented in R using the fastml[24] framework, which enforces guarded resampling and leakage-safe workflows, together with glmnet[25] for penalized regression.

We employed a nested cross-validation design, consisting of an outer stratified 10-fold cross-validation loop for unbiased performance estimation and an inner repeated 10-fold cross-validation (5 repeats) loop for hyperparameter tuning. Within each outer training fold, a grid search across α (0-1, step 0.05) and λ values was conducted to identify the optimal penalty parameters. This ensured that DEG selection, preprocessing, and hyperparameter optimization were confined entirely to training data in each resampling iteration. The optimal model consistently corresponded to $\alpha = 1$, indicating pure LASSO regularization.

Model discrimination was evaluated using ROC curves and area under the curves (AUCs) computed on held-out outer test folds using the pROC package[26]. Overall cross-validated performance was summarized by pooling predictions from all outer test folds, and 95% confidence intervals (CIs) were obtained using DeLong's method. The optimal decision threshold was determined by maximizing Youden's J statistic, from which sensitivity, specificity, accuracy, and predictive values were calculated.

To assess biomarker stability, we performed 1,000 bootstrap resamples of the discovery dataset, refitting the final leakage-corrected LASSO model for each resample using the optimal α and λ. The frequency with which each gene retained a nonzero coefficient was recorded to quantify selection stability. Features appearing in more than 50% of resamples were considered robust predictors. This stability analysis was conducted independently of cross-validation to characterize feature robustness rather than predictive performance. All modeling procedures adhered to TRIPOD guidelines[27], with explicit safeguards against data leakage at all stages of feature selection, model tuning, and evaluation.

Statistical Analysis

All data processing and analyses were performed in R (version 4.4.2). Throughout the analysis, reproducibility and transparency were maintained by setting random seeds for cross-validation (to ensure consistent partitioning) and by documenting all code steps. All statistical tests were two-tailed, and a significance level of 0.05 was used unless otherwise specified. The Methods were written adhering to STROBE[28] and REMARK[29] guidelines for observational transcriptomic analyses and biomarker evaluations, ensuring that the study can be replicated and built upon by other researchers.

RESULTS

Unsupervised Clustering of Samples

PCA of the discovery data (*GSE208194*) revealed distinct clustering of CAD patients versus controls. In the Figure 1A, PC1 (32 %) and PC2 (15 %) together capture roughly 47 % of the total variance. Most CAD samples cluster on the negative side of PC1, whereas control samples are enriched on the positive side, producing a clear but not complete group separation. The overlap near the PC1 origin indicates some heterogeneity within both cohorts, yet the overall pattern supports the existence of a disease-associated blood-transcriptomic signature. Similarly, a Uniform Manifold Approximation and Projection (UMAP) analysis reinforced the clustering pattern (Figure 1B). Most CAD samples project to the lower-left sector of the two-dimensional map, whereas control samples are enriched in the upper-right, producing two loose aggregations along the UMAP-1 axis. Although a noticeable subset of samples from each group intermingle in the central region, the overall distribution points to systematic transcriptomic differences between CAD and controls, justifying downstream differential-expression analyses.

Differentially Expressed mRNAs

Using limma moderated analysis, we identified a robust set of genes that were significantly differentially expressed between CAD patients and controls (Table S1). In total, 182 transcripts met the significance criteria ($|\log^{FC}| \geq 1$ and p-adjusted <0.05), of which 177 were protein-coding mRNAs and 5 were lncRNAs. Among these 182 DEGs, the majority (approximately 91%) showed lower expression in CAD relative to controls, while about 9% were up-regulated in CAD. In other words, 166 genes were down-regulated and 16 were up-regulated in CAD patients (Figure 2). This overall trend of suppressed gene expression in CAD blood is consistent with previous transcriptomic studies of CAD patients[7].

In our transcriptome analysis of CAD samples, the most strongly upregulated transcripts were mitochondrial-encoded oxidative phosphorylation genes. For example, *MT-ND1* (NADH dehydrogenase subunit 1) was markedly induced ($\log^{FC} = 1.712$, p-adjusted <0.001), as were *MT-ND6* ($\log^{FC} = 1.661$, p-adjusted <0.001), *MT-ND5* ($\log^{FC} = 1.624$, p-adjusted <0.001) and *MT-ND3*

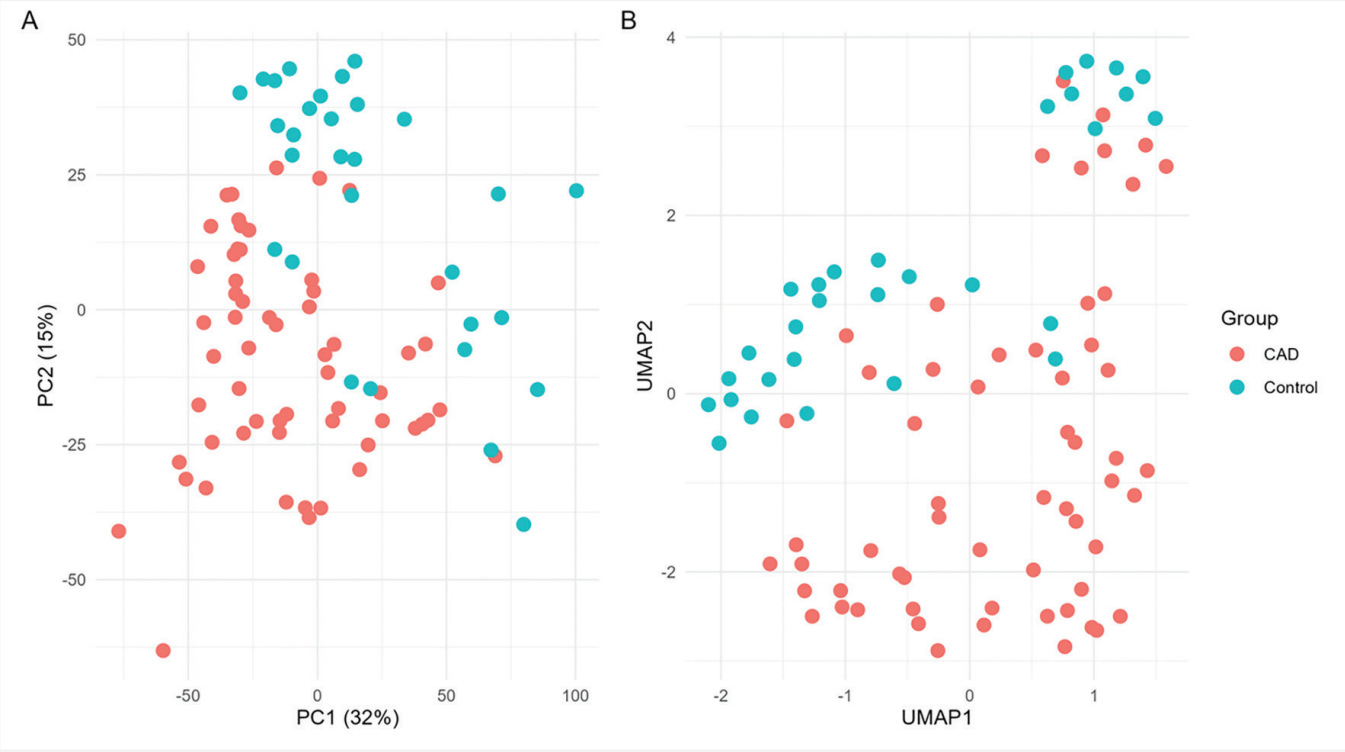

Figure 1. Unsupervised dimensionality-reduction of whole-blood transcriptomes from CAD patients and healthy controls (A) PCA, (B) UMAP
CAD: Coronary artery disease, PCA: Principal-component analysis, UMAP: Uniform Manifold Approximation and Projection

(log^{FC} =1.604, p-adjusted <0.001). These genes encode core subunits of mitochondrial Complex I (NADH dehydrogenase)[6]. Complex I is an essential component of the electron transport chain for adenosine triphosphate (ATP) production, and its dysfunction is known to increase mitochondrial reactive oxygen species and trigger inflammation and vascular remodeling in atherosclerotic cardiovascular disease[30]. Thus, the coordinated upregulation of multiple Complex I subunit genes in CAD samples may reflect altered mitochondrial activity. In support of this, other mitochondrial transcripts (e.g. *MT-CYB* for Complex III and *MT-CO2/MT-CO3* for Complex IV) were also modestly elevated. Overall, these results suggest that CAD is associated with enhanced expression of respiratory chain components, potentially as a compensatory response or in relation to increased oxidative stress in disease.

Conversely, several genes involved in mitochondrial biogenesis and function were among the most downregulated. Notably, *NDUFAF2* (a Complex I assembly factor) was significantly decreased (log^{FC} =-1.275, p-adjusted <0.001), as was *COA6* (a cytochrome c oxidase/Complex IV assembly factor; log^{FC} =-1.325, p-adjusted <0.001) and *TOMM5* (a subunit of the translocase of outer mitochondrial membrane; log^{FC} =-1.231, p-adjusted <0.001). Loss of *NDUFAF2* is known to impair Complex I assembly and cause mitochondrial dysfunction with increased oxidative stress[31], which can exacerbate endothelial injury and inflammation. Reduced *COA6* levels would likely hinder Complex IV maturation, further compromising electron transport. In addition, the endogenous retroviral envelope gene *ERV3-1* was strongly suppressed (log^{FC} =-1.485, p-adjusted <0.001); *ERV3* has been implicated in immune regulation (including autoimmunity)[32], so its downregulation might reflect altered innate immune signaling in CAD. Together, the downregulated genes indicate a pattern of mitochondrial compromise, with disrupted assembly of multiple respiratory complexes and import machinery, which is consistent with theories that mitochondrial dysfunction and resultant reactive oxygen species contribute to CAD pathogenesis[30,31].

Differentially Expressed lncRNAs

In addition to mRNAs, five lncRNAs showed significant differential expression between CAD and controls ($|log^{FC}| \geq 1$, FDR <0.05): *LINC03031*, *CCDC26*, *MAST4-AS1*, *MIR663AHG*, and *PCED1B-AS1*. *LINC03031* was upregulated (log^{FC} =1.182, p<0.001) and *MAST4-AS1* downregulated (log^{FC} =-1.361, p<0.001), representing novel, previously unreported candidates in cardiovascular disease. *CCDC26* was upregulated (log^{FC} =1.094, FDR <0.001) and has been identified as an oncogenic lncRNA[33] and one of the most elevated plasma lncRNAs in CAD[34] possibly

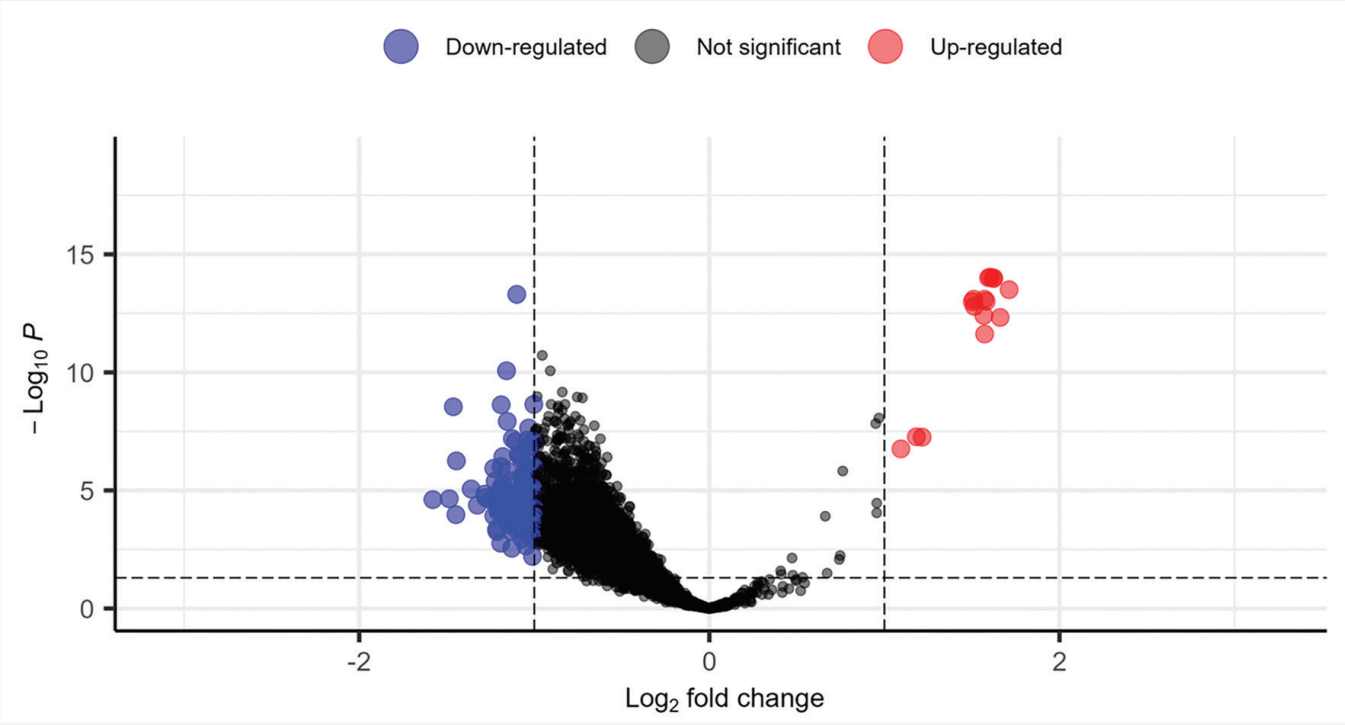

Figure 2. Volcano plot summarizing DEGs between CAD and control groups. The plot highlights the significant DEGs (|logFC | ≥1 and p-adjusted <0.05) in color, with the horizontal axis showing logFC (CAD/Control) and the vertical axis showing -log10(P). Notably, the distribution of points underscores the predominance of down-regulated genes in CAD (left side) compared to up-regulated genes (right side)

DEGs: Differentially expressed genes, CAD: Coronary artery disease

influencing vascular proliferation and inflammation. *PCED1B-AS1* was downregulated (log^{FC} =-1.072, p<0.001); known to promote cell proliferation via MAPK signaling[35], its reduction here may reflect suppressed growth signaling. *MIR663AHG* was also downregulated (log^{FC} =-1.579, p<0.001). As the host gene of miR-663a, which exerts anti-inflammatory effects[36], *MIR663AHG* repression may reduce vascular protection and favor pro-atherogenic signaling.[37] Collectively, these lncRNAs, particularly upregulated *CCDC26* and downregulated *PCED1B-AS1* and *MIR663AHG*, highlight biological pathways related to proliferation, hypoxia, and inflammation. Their stability and detectability in blood suggest value as noninvasive biomarkers for CAD detection, risk stratification, and therapy monitoring. Further validation and mechanistic studies are warranted to clarify their functional roles in atherosclerosis.

Pathway- and Gene Ontology Enrichment of DEGs

To identify biological pathways altered in CAD, we performed KEGG and GO enrichment analyses on DEGs (Figure 3). Seventeen KEGG pathways passed the 5% FDR threshold (Figure 4A). The top hit was "Ribosome" (hsa03010; 32/102 DEGs, p<0.001), followed by "coronavirus disease-19" and "thermogenesis," both sharing ribosomal genes. Mitochondrial pathways, including oxidative phosphorylation (15/102 genes, p<0.001) and diabetic cardiomyopathy, were also enriched, indicating up-regulated respiratory-chain activity (e.g., *NDUFA9, ATP5F1B*). Nucleotide excision repair, spliceosome, and RNA polymerase modules showed similar enrichment, suggesting intensified nucleic-acid maintenance. GO, BP confirmed enrichment in cytoplasmic translation, ribosome biogenesis, oxidative phosphorylation, and ATP synthesis (all p<0.001). p53-mediated apoptotic signaling further suggested DNA-damage responses. Cellular-component terms were dominated by ribosomal subunits and mitochondrial complexes I, IV, and ATP synthase, while spliceosomal complexes (U1, U2, U4/U6-U5) pointed to enhanced post-transcriptional activity. The top molecular-function term was "structural constituent of ribosome" (34/157 DEGs, p<0.001), followed by transporter and RNA-binding activities. These enrichments indicate a metabolic-inflammatory shift in CAD blood, characterized by heightened protein synthesis, mitochondrial energy production, and RNA processing, consistent with chronic immune activation and oxidative stress in atherosclerosis. Ribosomal and mitochondrial pathways are thereby highlighted as potential therapeutic targets.

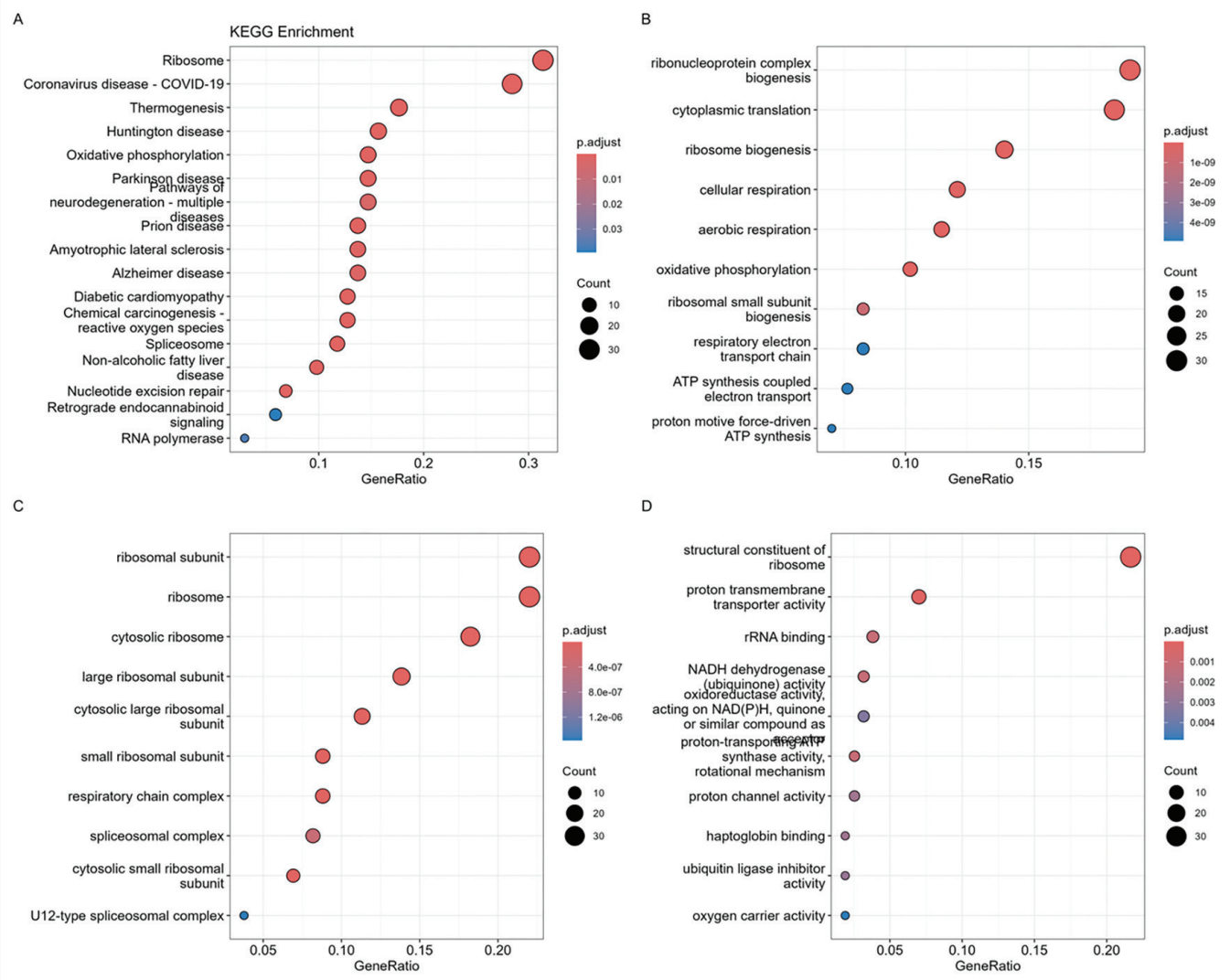

Figure 3. Functional enrichment of CAD-associated transcripts. Dot-plots display the top enriched terms for A) KEGG pathways, B) Gene ontology-biological process, C) Gene ontology-cellular component and D) Gene ontology-molecular function

CAD: Coronary artery disease, KEGG: Kyoto Encyclopedia of Genes and Genomes, ATP: Adenosine triphosphate

Validation Phase Results

In the validation phase, 85 genes were confirmed as differentially expressed, supporting the discovery-phase findings (Table S2). Effect-size concordance across platforms was quantified by correlating log^{FC} estimates between the discovery RNA-seq and validation microarray cohorts. For the validated genes (n=85; p-adjusted <0.05 in the validation cohort with concordant direction), effect sizes were strongly correlated (r=0.671; 95% CI 0.535-0.774; p<0.001). In contrast, when considering the full set of discovery DEGs represented on the microarray platform (n=1116), cross-platform correlation was weak and not statistically significant (r=0.0368; 95% CI -0.0219 to 0.0953; p=0.219), indicating that reproducible effect-size agreement is concentrated within the validated subset rather than across all discovery DEGs.

Eighty-three encoded proteins, while two were lncRNAs (*LINC03031* and *PCED1B-AS1*). Validation fold-changes were attenuated relative to discovery (median | log^{FC} | =0.39 *vs.* 1.07), a pattern consistent with cross-cohort and cross-platform replication. Most genes retained high statistical significance: 71 of 85 (84%) had adjusted $p<10^{-3}$, 47 (55%) $<10^{-6}$, 32 (38%) $<10^{-9}$, and 19 (22%) $<10^{-12}$, demonstrating robust reproducibility. *LINC03031* showed stronger up-regulation (log^{FC} = +1.385 vs +1.182), whereas *SRSF10*, *ARPC4-TTLL3*, and *RPS27* reproduced marked down-regulation. Other highly significant transcripts (*GNLY, CNOT6L, RPL7, TOMM5*) reinforced pathways related to ribosome biogenesis, mitochondrial function, and immune activation. Overall, gene-level changes were directionally consistent and statistically robust, though quantitative effect sizes diminished across cohorts.

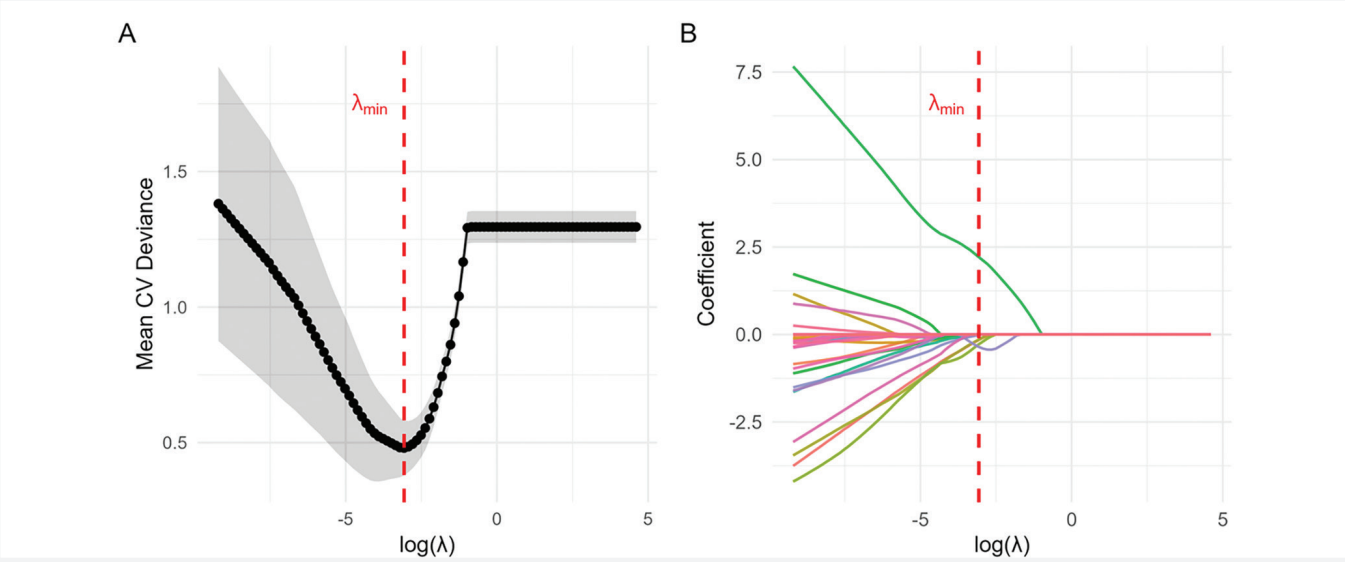

Figure 4. Regularized logistic regression model performance and coefficient profiles. (A) Cross-validation MSE plotted against the logarithm of the regularization parameter log(lambda). (B) Coefficient trajectories across the range of log(lambda) values MSE: Mean squared error, KEGG: Kyoto Encyclopedia of Genes and Genomes, GO: Gene ontology, ATP: Adenosine triphosphate, NADH: Nicotinamide adenine dinucleotide + hydrogen, COVID-19: Coronavirus disease-2019

Predictive Model Performance

The final predictive model was based on regularized logistic regression with LASSO penalization. Using leakage-controlled evaluation, the pooled cross-validated ROC AUC from the outer cross-validation loop was 0.975 (95% CI: 0.942-1.000), indicating strong discrimination under strict resampling conditions. This estimate was derived exclusively from held-out test folds, ensuring unbiased performance assessment.

When applied to the independent validation cohort (*GSE113079*), the model achieved an ROC AUC of 0.820 (95% CI: 0.737-0.906) (Figure 5), demonstrating preservation of predictive performance across platforms and biological compartments. At the Youden-optimized threshold, sensitivity =0.882, specificity =0.688, accuracy =0.823, balanced accuracy =0.785, positive predictive value =0.845, and negative predictive value =0.750.

Only two genes, *NEUROD2* and *RPS27*, retained non-zero coefficients in the final model (Figure 4B). Bootstrap stability analysis (1,000 resamples) showed consistent selection of *NEUROD2* in 99.9% and *RPS27* in 51.0% of runs, exceeding the predefined 50% stability threshold (Figure 6). Other transcripts appeared in fewer than 35% of resamples. These results indicate that predictive performance is driven by a minimal, stable two-gene signature rather than diffuse multigene effects.

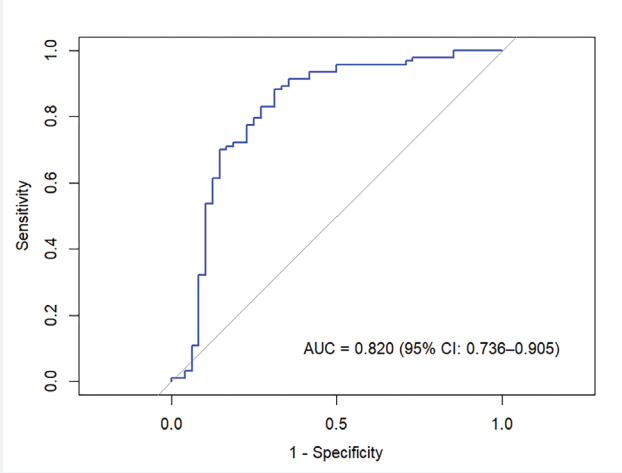

Figure 5. Receiver-operating-characteristic curve of the two-gene LASSO classifier in the validation cohort

AUC: Area under the curve, CI: Confidence interval

DISCUSSION

This study identified distinct blood transcriptomic signatures associated with CAD. Differential-expression analysis revealed coordinated dysregulation of mRNAs and lncRNAs, with enrichment in ribosomal biogenesis, oxidative phosphorylation, and inflammatory-stress pathways (p53, TNF/NF-κB). A concise

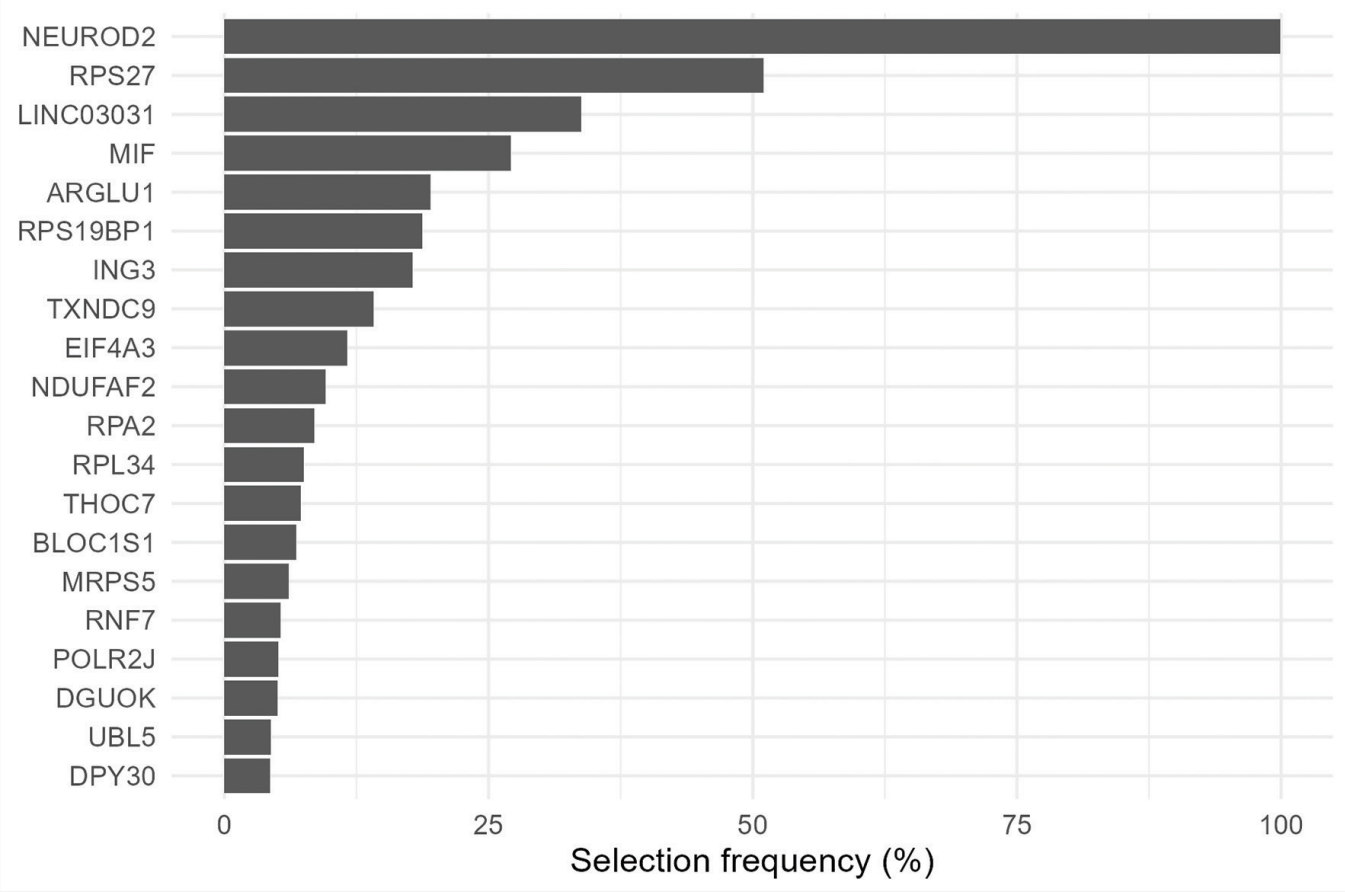

Figure 6. Top 20 features ranked by selection frequency across 1000 bootstrap resamples using the regularized logistic regression model at optimal alpha and lambda

AUC: Area under the curve, CI: Confidence interval

two-gene signature (*NEUROD2*, *RPS27*) accurately discriminated patients from controls, while *LINC03031* and *PCED1B-AS1* emerged as reproducible lncRNA markers, all validated across platforms. *NEUROD2*, previously recognized as a neuronal transcription factor, was recently found elevated in stable CAD[34] and implicated in macrophage inflammation via protein kinase D and *NLRP3/NF-κB*[38]. *RPS27*, a ribosomal protein, modulates p53 and NF-κB signaling[39,40], indicating its role in immune activation. *LINC03031*, largely uncharacterized, was strongly upregulated, representing a new CAD-linked transcript, while *PCED1B-AS1*, known from cancer biology, was downregulated, potentially reflecting suppressed hypoxia or growth-factor signaling[41]. These findings extend prior transcriptomic diagnostic efforts such as Corus CAD[11], PREDICT[42], and later multi-gene models[34,43]. In contrast to these larger gene panels, which typically comprise 12-23 transcripts and were optimized within narrowly defined clinical populations, the present study demonstrates that a highly compact two-gene signature can retain robust discriminatory performance when externally validated across cohorts, platforms, and biological compartments. Our results confirm *NEUROD2* as a reproducible CAD marker and highlight ribosomal-stress involvement through *RPS27*, defining a minimal two-gene panel with comparable diagnostic accuracy. The extreme parsimony of this model reduces model complexity, limits overfitting risk, and facilitates assay standardization, supporting feasibility for low-cost, qPCR-based implementation. Clinically, this compact panel could complement existing tests by offering a low-cost, qPCR-based assay for early or atypical CAD detection. Because it captures inflammatory and stress-response activity, it may yield information beyond lipid or imaging data.

An important aspect of this study is the biological distinction between the discovery and validation samples. Plasma cfRNA represents a heterogeneous mixture of extracellular transcripts released from multiple tissues, including vascular, cardiac, and immune-related cell types, whereas PBMC transcriptomes predominantly reflect gene expression within circulating immune cells. Prior work has demonstrated that cfRNA captures broad, systemic disease-associated transcriptional

programs rather than signals confined to a single tissue or cell type[44]. Accordingly, replication of key signals across biologically distinct blood-derived RNA sources supports the robustness and generalizability of the identified transcriptomic patterns, while necessarily limiting precise inference about tissue or cellular origin. This reduced spatial resolution limits biological specificity for localized processes, such as coronary plaque-restricted pathology. The observed signals are therefore most consistent with systemic transcriptional programs that are hypothesis-consistent with known features of CAD, including chronic vascular inflammation, immune activation, mitochondrial stress, and altered protein synthesis, rather than direct evidence of plaque-localized mechanisms.

At the same time, the biological nature of the validated signals warrants careful interpretation. Several prominent components of the signature, particularly ribosomal and mitochondrial transcripts, are consistent with generalized inflammatory and metabolic stress responses rather than being uniquely specific to CAD. Ribosomal proteins such as *RPS27* have been shown to participate in immune activation and stress-related signaling pathways across diverse disease contexts, reflecting broad host response programs rather than tissue-restricted pathology[45]. Accordingly, the presence of such transcripts suggests engagement of systemic immune and stress-associated processes. The validated non-coding transcripts are therefore most plausibly interpreted as reflecting regulatory programs related to cellular stress adaptation rather than disease-specific markers. These features are hypothesis-consistent with established models of atheroinflammation, in which sustained immune activation, mitochondrial dysfunction, and translational stress contribute to vascular injury and disease progression, but they do not provide direct evidence of plaque-localized or CAD-specific molecular mechanisms.

These considerations indicate that the identified transcriptomic signature is most appropriately interpreted as reflecting systemic inflammatory and stress-related processes rather than a plaque-specific molecular fingerprint or a disease-specific marker uniquely distinguishing CAD from other inflammatory conditions. Prior work has shown that plasma cfRNA captures broad, multi-tissue transcriptional programs and has limited spatial resolution with respect to tissue and lesion specificity[44]. Accordingly, the present findings are hypothesis-consistent with an association between the observed signature and overall atheroinflammatory burden in CAD, but they do not provide direct evidence of plaque-localized pathology or anatomic coronary stenosis. From a clinical perspective, these features suggest that such a signature, if validated in prospective studies, may be more appropriately considered for applications such as blood-based screening or risk stratification rather than as a definitive diagnostic of obstructive coronary disease. Potential use cases may include identifying individuals with elevated systemic inflammatory activity in whom further cardiovascular evaluation may be warranted, or complementing established risk factors and imaging modalities. These proposed applications remain speculative and require dedicated clinical validation beyond the scope of the present study.

Future validation efforts should focus on three key areas: (i) prospective evaluation in larger and more diverse populations to establish generalizability and reference ranges; (ii) assessment of specificity relative to other cardiovascular and systemic inflammatory diseases; and (iii) testing integrated models combining transcriptomic markers with established clinical predictors to quantify incremental diagnostic and prognostic value. Mechanistic studies should define their roles in vascular inflammation and remodeling, while integrated multi-omic and prospective evaluations will be critical for clinical translation.

Study Limitations

This study has several limitations. Although the findings were validated across independent cohorts and platforms, sample sizes were modest and relatively homogeneous, warranting confirmation in larger and more diverse populations. The use of biologically distinct compartments, plasma cfRNA in discovery and PBMCs in validation, strengthens robustness but complicates interpretation of tissue specificity and cellular origin. Although raw RNA-seq reads for the discovery cohort are available in the SRA, a gene-level raw count matrix was not provided as part of the GEO Series record; reprocessing FASTQ files to generate counts would constitute a new primary analysis and was beyond the scope of this secondary study. Gene-expression analyses were not adjusted for clinical covariates, leaving potential residual confounding. In addition, the identified transcriptomic signals may reflect general inflammatory or metabolic stress rather than CAD-specific processes, and specificity relative to other cardiovascular or systemic inflammatory diseases remains untested. Finally, the cross-sectional design precludes causal inference, and prospective clinical validation is required to establish diagnostic utility and define optimal clinical applications.

CONCLUSION

We identified and cross-validated a concise two-gene blood transcriptomic signature, *NEUROD2* and *RPS27*, that distinguishes patients with CAD from controls with robust discriminatory performance (nested cross-validated AUC=0.975; external AUC=0.820) across independent cohorts and platforms. This mRNA pair reflects ribosome- and mitochondria-associated inflammatory and metabolic stress pathways highlighted by enrichment analyses, complementing

established biomarkers rather than replacing them. Although *LINC03031* and *PCED1B-AS1* were not retained in the final predictive model, their reproducible dysregulation supports the broader relevance of lncRNAs in CAD-associated transcriptional alterations. Collectively, these findings provide a biologically coherent and technically validated framework for the further development of a blood-based transcriptomic assay, pending confirmation in larger, ethnically diverse populations and prospective studies to define clinical utility and application.

Ethics

Ethical Committee Approval: Ethics approval was not required because this study was based exclusively on publicly available de-identified secondary datasets from the GEO database. No new human participants were recruited, and no identifiable personal or clinical data were accessed.

Informed Consent: This study was a retrospective bioinformatics analysis of two publicly available GEO transcriptomic datasets on CAD.

Data Availability Statement

The data that support the findings of this study are publicly available in the NCBI GEO database. The associated GSE numbers are as follows: *GSE208194* and *GSE113079*. These datasets can be accessed directly via the following link: NCBI GEO database. For any additional information or specific data requests, please contact the corresponding author.

Data Deposition

All codes are available on GitHub repository: https://github.com/selcukorkmaz/blood-transcriptome-cad

Footnotes

Authorship Contributions

Concept: B.E.Y., S.K., Design: B.E.Y., S.K., Data Collection or Processing: S.K., Analysis or Interpretation: B.E.Y., S.K., Literature Search: B.E.Y., S.K., Writing: B.E.Y., S.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary File: https://d2v96fxpocvxx.cloudfront.net/37eae217-e8b5-4f55-976f-35df98003e83/content-images/7a123a81-927b-4901-9cc5-91fbcce6b3fa.pdf

REFERENCES

1. Di Cesare M, Perel P, Taylor S, Kabudula C, Bixby H, Gaziano TA, et al. The Heart of the World. Glob Heart. 2024;19:11.
2. Mensah GA, Fuster V, Murray CJL, Roth GA; Global burden of cardiovascular diseases and risks collaborators. Global burden of cardiovascular diseases and risks, 1990-2022. J Am Coll Cardiol. 2023;82:2350-473.
3. Zaman S, Wasfy JH, Kapil V, Ziaeian B, Parsonage WA, Sriswasdi S, et al. The lancet commission on rethinking coronary artery disease: moving from ischaemia to atheroma. Lancet. 2025;405:1264-312.
4. Zhang Z, Salisbury D, Sallam T. Long noncoding RNAs in Atherosclerosis: JACC review topic of the week. J Am Coll Cardiol. 2018;72:2380-90.
5. Li X, Zhang Y, Ding Z, Chen Y, Wang W. LncRNA H19: A novel biomarker in cardiovascular disease. Acta Cardiol Sin. 2024;40:172-81.
6. Rai V. Current and future role of biomarkers in the monitoring and prognosis of coronary artery disease. Future Cardiol. 2025;21:331-3.
7. McCaffrey TA, Toma I, Yang Z, Katz R, Reiner J, Mazhari R, et al. RNAseq profiling of blood from patients with coronary artery disease: signature of a T cell imbalance. J Mol Cell Cardiol Plus. 2023;4:100033.
8. Siemelink MA, Zeller T. Biomarkers of coronary artery disease: the promise of the transcriptome. Curr Cardiol Rep. 2014;16:513.
9. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC Med Genomics. 2011;4:26.
10. Voora D, Coles A, Lee KL, Hoffmann U, Wingrove JA, Rhees B, et al. An age- and sex-specific gene expression score is associated with revascularization and coronary artery disease: insights from the prospective multicenter imaging study for evaluation of chest pain (PROMISE) trial. Am Heart J. 2017;184:133-40.
11. Rosenberg S, Elashoff MR, Lieu HD, Brown BO, Kraus WE, Schwartz RS, et al. Whole blood gene expression testing for coronary artery disease in nondiabetic patients: major adverse cardiovascular events and interventions in the PREDICT trial. J Cardiovasc Transl Res. 2012;5:366-74.
12. Zhang YH, Pan X, Zeng T, Chen L, Huang T, Cai YD. Identifying the RNA signatures of coronary artery disease from combined lncRNA and mRNA expression profiles. Genomics. 2020;112:4945-58.
13. Kessler T, Schunkert H. Coronary artery disease genetics enlightened by genome-wide association studies. JACC Basic Transl Sci. 2021;6:610-23.
14. Huang J, Li M, Li J, Liang B, Chen Z, Yang J, et al. LncRNA H19 rs4929984 variant is associated with coronary artery disease susceptibility in han chinese female population. Biochem Genet. 2021;59:1359-80.
15. Wang XM, Li XM, Song N, Zhai H, Gao XM, Yang YN. Long non-coding RNAs H19, MALAT1 and MIAT as potential novel biomarkers for diagnosis of acute myocardial infarction. Biomed Pharmacother. 2019;118:109208.
16. Mu J, Chen C, Ren Z, Liu F, Gu X, Sun J, et al. Multicenter validation of lncRNA and target mRNA diagnostic and prognostic biomarkers of acute ischemic stroke from peripheral blood leukocytes. J Am Heart Assoc. 2024;13:e034764.
17. Elashoff MR, Nuttall R, Beineke P, Doctolero MH, Dickson M, Johnson AM, et al. Identification of factors contributing to variability in a blood-based gene expression test. PLoS One. 2012;7:e40068.
18. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846-7.
19. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4:1521.
20. Li L, Wang L, Li H, Han X, Chen S, Yang B, et al. Characterization of LncRNA expression profile and identification of novel LncRNA biomarkers to diagnose coronary artery disease. Atherosclerosis. 2018;275:359-67.
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
22. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141.
23. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, et al. Using clusterProfiler to characterize multiomics data. Nat Protoc. 2024;19:3292-320.

24. Korkmaz S, Goksuluk D, Karaismailoglu E. fastml: Guarded resampling workflows for safe and automated machine learning in R. R package version 0.7.7. 2026. Available from: https://CRAN.R-project.org/package=fastml

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

26. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.

28. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke J, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Int J Surg. 2014;12:1495-9.

29. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. PLoS Med. 2012;9:e1001216.

30. Chen Y, Yang M, Huang W, Chen W, Zhao Y, Schulte ML, et al. Mitochondrial metabolic reprogramming by CD36 signaling drives macrophage inflammatory responses. Circ Res. 2019;125:1087-102.

31. Park S, Trujillo-Hernandez JA, Levine RL. Ndufaf2, a protein in mitochondrial complex I, interacts in vivo with methionine sulfoxide reductases. Redox Rep. 2023;28:2168635.

32. Bustamante Rivera YY, Brütting C, Schmidt C, Volkmer I, Staege MS. Endogenous retrovirus 3 - history, physiology, and pathology. Front Microbiol. 2018;8:2691.

33. Wang S, Hui Y, Li X, Jia Q. Silencing of lncRNA CCDC26 restrains the growth and migration of glioma cells in vitro and in vivo via targeting miR-203. Oncol Res. 2018;26:1143-54.

34. Ward Z, Schmeier S, Pearson J, Cameron VA, Frampton CM, Troughton RW, et al. Identifying candidate circulating RNA markers for coronary artery disease by deep RNA-sequencing in human plasma. Cells. 2022;11:3191.

35. Cao J, Yang Y, Duan B, Zhang H, Xu Q, Han J, et al. LncRNA PCED1B-AS1 mediates miR-3681-3p/MAP2K7 axis to promote metastasis, invasion and EMT in gastric cancer. Biol Direct. 2024;19:34.

36. Yuan H, Ren Q, Du Y, Ma Y, Gu L, Zhou J, et al. LncRNA miR663AHG represses the development of colon cancer in a miR663a-dependent manner. Cell Death Discov. 2023;9:220.

37. Arencibia A, Lanas F, Salazar LA. Long non-coding RNAs might regulate phenotypic switch of vascular smooth muscle cells acting as ceRNA: implications for in-Stent restenosis. Int J Mol Sci. 2022;23:3074.

38. Tan X, Yan C, Zou G, Jing R. Neurogenic differentiation 2 promotes inflammatory activation of macrophages in doxorubicin-induced myocarditis via regulating protein kinase D. BMC Cardiovasc Disord. 2025;25:195.

39. Feng J, Li Y, Wang C, Wang Y, Wan Y, Zheng M, et al. Peripheral blood transcriptomic analysis identifies potential inflammation and immune signatures for central retinal artery occlusion. Sci Rep. 2024;14:7398.

40. DeGroat W, Abdelhalim H, Peker E, Sheth N, Narayanan R, Zeeshan S, et al. Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases. Sci Rep. 2024;14:26503.

41. Yao Z, Zhang Q, Guo F, Guo S, Yang B, Liu B, et al. Long Noncoding RNA PCED1B-AS1 promotes the warburg effect and tumorigenesis by upregulating HIF-1α in glioblastoma. Cell Transplant. 2020;29:963689720906777.

42. Wang XB, Cui NH, Liu X, Ming L. Identification of a blood-based 12-gene signature that predicts the severity of coronary artery stenosis: an integrative approach based on gene network construction, support vector machine algorithm, and multi-cohort validation. Atherosclerosis. 2019;291:34-43.

43. Xing Y, Lin X. Transcriptome associated with single-cell analysis reveal the role of S-palmitoylation in coronary artery disease. Sci Rep. 2025;15:15144.

44. Vorperian SK, Moufarrej MN; Tabula Sapiens Consortium; Quake SR. Cell types of origin of the cell-free transcriptome. Nat Biotechnol. 2022;40:855-61.

45. Diao MQ, Li C, Xu JD, Zhao XF, Wang JX. RPS27, a sORF-encoded polypeptide, functions antivirally by activating the NF-κB pathway and interacting with viral envelope proteins in shrimp. Front Immunol. 2019;10:2763.