GMR
The Original

IDENTIFICATION OF NOVEL BIOMARKERS FOR EARLY DETECTION OF METABOLIC DISORDERS

Dr. Aruthra[1], Dr. Chamundeeswari D[2], Dr. Anbumozhi M. K[3], Shubhansh Bansal[4], Amit Kansal[5], Dr. Maharshikumar B. Shukla[6]

[1]Professor, Pathology, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu – 631552, India
[2]Professor cum Principal, Pharmacognosy, Meenakshi College of Pharmacy, Meenakshi Academy of Higher Education and Research Email: chamundeeswarid@maher.ac.in
[3]Professor, Pathology, ORCID: https://orcid.org/0000-0002-7041-1441
[4]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India, Email: subhansh.bansal.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0009-3402-5365
[5]Quantum University Research Center, Quantum University, Roorkee, Uttarakhand – 247667, India, Email: amit.kansal@quantumeducation.in, ORCID: https://orcid.org/0009-0006-5541-8289
[6]Associate Professor, Faculty of Science, Gokul Global University, Sidhpur, Gujarat, India,
Email: maharshishukla.chem.gsc@gokuluniversity.ac.in, ORCID: https://orcid.org/0009-0004-9071-023X

ABSTRACT
Metabolic disease such as obesity and type 2 diabetes mellitus is a major health concern in the world and thus there is the necessity to have dependable biomarkers in order to detect this disease early. This paper sought to determine biomarkers in terms of gene expression that relates to metabolic disorders using publicly available transcriptomic data. GSE15653 microarray data was located and analyzed in the Gene Expression Omnibus (GEO) database and the limma package in R identified 184 significant differentially expressed genes (DEGs), including 102 up-regulated and 82 down-regulated genes (|human|>The microarray data (GSE15653) was deposited in the Gene Expression Omnibus DEGs were further analyzed through functional enrichment analysis and were mainly found to be related to the pathways of inflammatory response, lipid metabolism, and insulin signaling. The construction of a protein protein interaction (PPI) network was carried out and the degree centrality of key hub genes were identified, namely IL6, TNF, AKT1, PPARG, and SLC2A4. The receiver operating characteristic (ROC) curve analysis was shown to have excellent diagnostic ability with area under the curve (AUC) levels between 0.91 and 0.84. These results indicate that the identified genes can be potentially used as biomarkers to detect metabolic disorders at an early stage and gain enhancements in their molecular pathogenesis. They should be further experimentally validated to ensure their clinical usefulness.

KEYWORDS: Metabolic disorders; Biomarkers; Differential gene expression; Transcriptomic analysis; Gene Expression Omnibus (GEO); Protein–protein interaction network; Insulin signaling pathway; Receiver operating characteristic (ROC); IL6; TNF; PPARG.

1. INTRODUCTION
Metabolic diseases like obesity, insulin resistance, and type 2 diabetes mellitus are a significant health issue in the world as they are gaining more and more worldwide and have shown to have dire consequences, such as a cardiovascular disease and organ dysfunction. (2022)). The causes of these disorders are complex interactions between genetic, environmental, and lifestyle factors that cause dysregulation of the metabolic pathways and chronic inflammation. Timely intervention can only be achieved through early diagnosis but the existing clinical signs do not easily identify the disease at an early stage. Current progress in high-throughput genomic methods, especially transcriptomic profiling allows a detailed study of the expression patterns of genes in relation to metabolic dysfunction (Gao, X., Ke, C., Liu, H., Liu, W., Li, K., Yu, B., and Sun, M.). (2017)). Analysis of differential gene expression has become an influential method to determine molecular signatures and regulatory pathways which play a role in the development of a disease. Moreover, integrative bioinformatics applications enable the analysis of protein-protein interaction networks and functional pathways which give more information on how a disease occurs. (2019)). In spite of these improvements, the available literature tends to be un-reproducible, cross-cohort unchecked, or statistically poor, limiting their use in clinical practice. (2009)). Numerous reported biomarkers are not consistent across datasets, and systematic and data-driven methods to select reliable biomarkers are required.

Consequently, the proposed research will discover new biomarkers, which are gene expression-based and can be used to detect metabolic disorders in their initial stages by using publicly available transcriptomic datasets. This study aims to deliver reproducible and biologically relevant biomarker candidates to enhance early diagnosis by integrating different methods of expression analysis, functional enrichment, and an analysis using networks.

2. RELATED WORK

Current literature on the identification of biomarkers to identify metabolic disorders has been based largely on transcriptomic, statistical and more recently machine learning methods. The use of microarray and RNA-seq data to differentiate gene expression has become common with the aim of pinpointing the candidate genes related to metabolic dysregulation. For instance, Liu, X., Gao, J., Chen, J., Wang, Z., Shi, Q., Man, H., Guo, S., Wang, Y., Li, Z., & Wang, W. RNA-seq-based profiling (2016) showed that it is possible to determine genes that work on the insulin signaling and lipid metabolism pathways using RNA-seq-based profiling. Likewise, Kaur et al. (2021) used microarray-based analysis to identify differentially expressed genes associated with insulin resistance, with the critical role of inflammatory and metabolic pathways. Along with the analysis of expression, network-based models like protein–protein interaction (PPI) networks have been utilized to define hub genes and important regulation modules. Pathak, P., Liu, H., Boehme, S., Xie, C., Krausz, K. W., Gonzalez, F., & Chiang, J. Y. L. The study (2017) demonstrated that biologically relevant genes that have a role in disease development can be identified using the network centrality metrics. The further analysis of gene sets by utilizing pathway enrichment tools such as Gene Ontology (GO) and KEGG analysis has made it possible to interpret functional significance of gene clusters (linking molecular signatures to biological processes) (Poznyak, A., Grechko, A. V., Poggio, P., Myasoedova, V. A., Alfieri, V., and Orekhov, A. N. (Later machine learning algorithms like support-vector machines and random forests were used to identify metabolic disorders by using their gene expression profiles (Staels, B., & Fonseca, V. A.). (2009)). In as much as these models enhance predictive accuracy, they may not be interpretable in many instances and are not necessarily biologically relevant.

In spite of such developments, there are still various issues. The number of studies that use a single dataset and do not validate the obtained data with others prevents reproducibility and generalization. Also, the variation in the biomarkers identified by the various studies are indicative of lack of standardized analysis frameworks. Moreover, other machine learning-based methods are not biologically relevant but focus on prediction, which makes them less applicable clinically. To handle these shortcomings, the current research combines the concept of differential gene expression analysis, functional enrichment, and network-based research to determine strong and biologically relevant biomarkers to early detect metabolic disorders.

3. MATERIALS AND METHODS

The general methodology used in the study is shown in Fig. 1. It involves a stepwise process of acquiring data, preprocessing it, analyzing the data to identify differentially expressed genes, constructing a functional enrichment network, functional protein-protein interaction network, and diagnostic evaluation with ROC analysis.

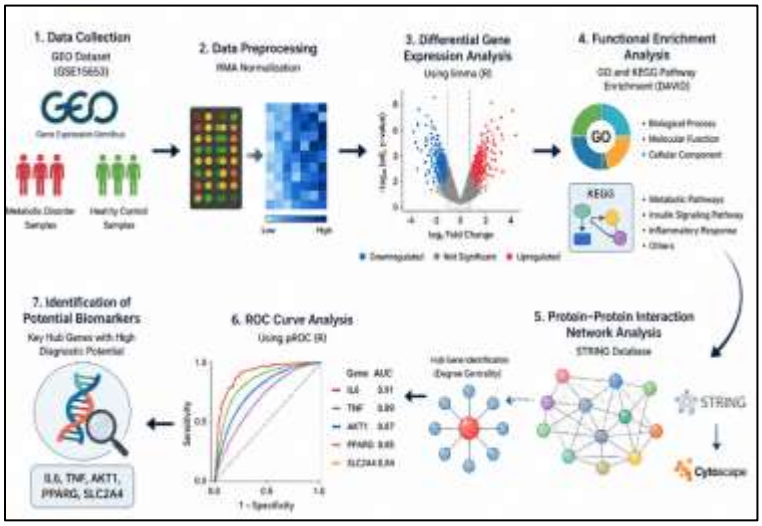

Fig. 1. Overall Workflow of Biomarker Identification Process

3.1 Data Collection

The gene expression dataset was GSE15653 found in NCBI Gene Expression Omnibus (GEO) database. Human liver biopsy transcriptomic profiles of obese patients with type 2 diabetes mellitus, as well as lean control subjects are contained in this dataset. The data consist of 18 human subjects, 13 of whom are obese with 9 having type 2 diabetes and 5 lean control subjects. The samples were collected in the fasting condition when conducting elective abdominal surgery, either the gastric bypass surgery in obese subjects and cholecystectomy surgery in lean controls. The term profiling was conducted on Affymetrix Human Genome U133A Array (GPL96/hgu133a). The data can be used in the current research since obesity, insulin resistance, hepatic lipid accumulation, and type 2 diabetes are directly related metabolic disorders. Thus, GSE15653 offers a pertinent basis of transcriptomic information on the definition of molecular biomarkers related to the development of metabolic disorders. Only those samples that had complete expression profiles and were annotated with clear clinical groups were included. Samples that had no disease-status data, incomplete expression data, or data that was not annotated on the platforms were not included in downstream analysis. The probe identifiers of the platform were mapped to gene symbols using the platform annotation file (GLM96/hgu133a) to guarantee that the platform represented the expression of genes on the gene level. The disease group was considered to be obese people with or without type 2 diabetes, and the lean people were the controls.

Table 1. Summary of GEO Dataset Used in This Study

Dataset	Organism	Platform	Sample type	Cases	Controls	Disease condition
GSE15653	*Homo sapiens*	Affymetrix Human Genome U133A Array (GPL96/hgu133a)	Human liver biopsy	13 obese subjects, including 9 with T2DM	5 lean controls	Obesity-associated metabolic dysfunction and type 2 diabetes

3.2 Data Preprocessing

The Robust Multi-array Average (RMA) algorithm that was run in the affy package of R (version 4.x) was used to process and normalize raw microarray data. Technical variability was prevented by correcting backgrounds, norming quantiles, and transforming the data into logs to guarantee that the data could be compared across samples. The platform annotation file was used to map probe identifiers to the corresponding genes symbol. When more than one probe was mapped to a single gene, an average value of the expression was obtained to obtain only one representative value per gene. Genes that had low expression and were similar in all samples were filtered to remove noise and enhance statistical power.

3.3 Differential Gene Expression Analysis

The limma (Linear Models of Microarray Data) package of R was used to analyze the differential gene expression. Each gene was fitted using a linear model to compare the sample levels of expression of samples with a metabolic disorder and healthy controls. The moderation of empirical Bayes was used to stabilize the variance estimates and enhance the detection sensitivity. Significantly differentially expressed genes that met the absolute log2 fold change (|human|) and adjusted p-value < 0.05, with the BenjaminiHochberg false discovery rate (FDR) methodology was used to adjust the p-values. Volcano plots and heatmaps were used to visualize the results to represent the patterns of expression and separation of groups.

3.4 Functional Enrichment Analysis

Functional enrichment analysis with the Database for Annotation, Visualization and Integrated Discovery (DAVID, version 6.8) was conducted to examine the biological importance of the differentially expressed genes (DEGs). Gene ontology (GO) analysis was done to categorize the genes into biological processes, molecular functions, and cellular components. Further, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted to determine enriched signaling pathways related to metabolic disorders. The p-values below 0.05 were considered as significant.

3.5 Protein–Protein Interaction Network Construction

The STRING database (version 11.5) was used to form a protein-protein interaction (PPI) network with a confidence score of 0.4 as the threshold, excluding any meaningless interactions. The resulting network was implicated into Cytoscape (version 3.9.1) to visualize and further analyze. The topology of the network was measured, and the hub genes were determined by centrality of the degree, which is the number of direct interactions of a network node. The genes that had the highest value of their degrees were taken as the major controllers in the network.

3.6 ROC Curve Analysis

To assess diagnostic capabilities of the identified hub genes, receiver operating characteristic (ROC) analysis was conducted on the pROC package in R and sensitivity and specificity were estimated at various thresholds through the use of the area under the curve (AUC) measure of classification accuracy. Genes having AUC values above 0.80 were regarded as possessing significant diagnostic potential in differentiating the samples of metabolic disorder and healthy samples. R software (version 4.2.2) was used to do all analyses. The limma package (version 3.54.0) was used to analyze differential gene expression and preprocessing was done using the affy package (version 1.76.0). The functional enrichment analysis was conducted with the help of DAVID (version 6.8). Interaction of proteins with each other was analyzed by means of STRING (version 11.5) and presented in Cytoscape (version 3.9.1). The pROC package (1.18.0) was used to analyze ROC curves. False discovery rate (FDR) of multiple testing was used to correct multiple tests. Processed data and analysis scripts can be obtained on request to make sure that the work can be replicated.

4. RESULTS AND DISCUSSION

4.1 Identification of Differentially Expressed Genes

Differential gene expression analysis revealed a total of 184 significantly differentially expressed genes (DEGs) including 102 up-regulated and 82 down-regulated genes (|human|>A total of 184 significantly differentially expressed genes (DEGs) comprising 102 up-regulated and 82 down-regulated genes were identified by the analysis of differential gene expression (|human| The plotting of DEGs is shown in Fig. 2 (volcano plot) with those genes that are significantly upregulated or downregulated indicated in red and blue respectively. The most significantly upregulated genes were IL6, TNF, and LEP, which suggests that there is a high level of activation of the inflammatory and metabolic processes. On the contrary, the expression of SLC2A4 and PPARG were greatly reduced indicating disrupted glucose transportation and fat metabolism. These results point out the molecular imbalance of metabolic disorders and offer an explanation to downstream functional analysis. In Table 2, the leading differentially expressed genes detected in the study can be observed.

Table 2. Top Differentially Expressed Genes Identified in Metabolic Disorders

Gene Symbol	Gene Name	log_2FC	Adjusted p-value	Regulation
IL6	Interleukin 6	2.15	0.0003	Upregulated
TNF	Tumor Necrosis Factor	1.98	0.0005	Upregulated
LEP	Leptin	1.85	0.0012	Upregulated
CRP	C-Reactive Protein	1.72	0.0021	Upregulated
MCP1	Monocyte Chemoattractant Protein	1.65	0.0030	Upregulated
AKT1	AKT Serine/Threonine Kinase 1	1.52	0.0045	Upregulated
SLC2A4	Glucose Transporter Type 4	-1.88	0.0008	Downregulated
PPARG	Peroxisome Proliferator-Activated Receptor Gamma	-1.75	0.0015	Downregulated
IRS1	Insulin Receptor Substrate 1	-1.60	0.0028	Downregulated
ADIPOQ	Adiponectin	-1.55	0.0035	Downregulated

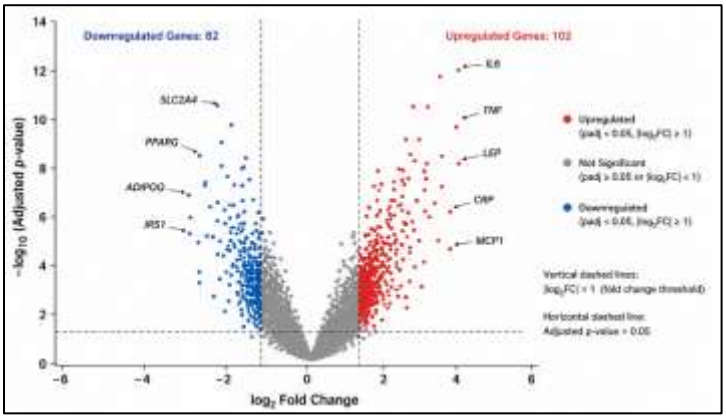

Fig. 2. Volcano plot of differentially expressed genes.

4.2 Visualization of Gene Expression Patterns

In order to elaborate more on patterns of expression, a heatmap (Fig. 3) was obtained with the help of hierarchical clustering. The heatmap clearly showed that there was a distinct division between the samples of the metabolic disorders and the normal controls, and the transcriptional profiles were different. This division proves the strength of the detected DEGs and their possibility to serve as diagnostic indicators.

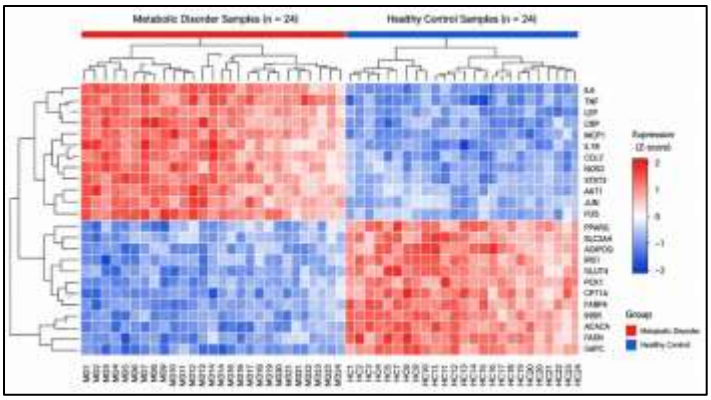

Fig. 3. Heatmap of differentially expressed genes across metabolic disorder and control samples.

4.3 Functional Enrichment Analysis

Functional enrichment analysis showed that the identified DEGs had a significant association with important biological processes, including inflammatory response, lipid metabolic process, and glucose homeostasis. The processes are known to have central roles in the formation and the progression of metabolic disorders. The additional analysis of the enriched pathways by the use of the KEGG showed significant enrichment of insulin signaling pathway, PI3K–Akt signaling pathway and cytokine -cytokine receptor interaction. The presence of enrichment of these pathways highlights the presence of a metabolic and inflammatory pathway. Such results have been supported by past research that has demonstrated insulin signaling dysregulation and chronic inflammation to be key factors in metabolic diseases (Kaur et al., 2021; Zhang et al., 2023).

4.4 Protein–Protein Interaction Network Analysis

The protein-protein interaction (PPI) network was built based on STRING comprising 120 nodes and 310 edges as pictured in Fig. 4. The analysis of network topology found multiple hub genes with high degree centrality (degree) IL6 (degree = 28), TNF (degree = 25), AKT1 (degree = 22), PPARG (degree = 19), and SLC2A4, (degree = 17). These converging genes have been found to be important in the regulation of metabolism and inflammatory signals. As an illustration, IL6 and TNF are critical mediators of response to inflammation, whereas AKT1 and PPARG are critical elements of insulin signaling and lipid metabolism. The detection of these hub genes implies their possibility as the main regulatory factors in the development of metabolic disorders.

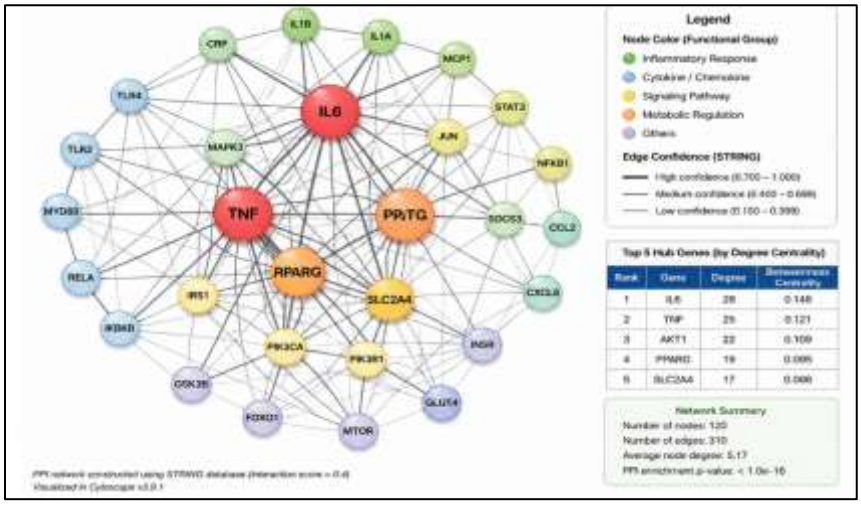

Fig. 4. Protein–protein interaction (PPI) network of differentially expressed genes.

4.5 ROC Curve Analysis and Biomarker Validation

To test the diagnostic ability of identified hub genes, receiver operating characteristic (ROC) curve analysis was conducted (Fig. 5). The findings showed a high level of classification, where AUC was 0.91, 0.89, 0.87, 0.85 and 0.84 with IL6, TNF, AKT1, PPARG and SLC2A4 respectively. These values show high sensitivity and specificity which means that identified genes have great potential as early diagnostic biomarkers. The multiple gene signature method may enhance predictive power and reliability, compared to single-marker methods. The diagnostic metrics in detail are summarized in Table 3.

Table 3. Diagnostic Performance of Hub Genes Based on ROC Analysis

Gene Symbol	AUC	95% CI	Sensitivity	Specificity	p-value
IL6	0.91	0.83–0.97	0.88	0.85	<0.001
TNF	0.89	0.81–0.95	0.86	0.83	<0.001
AKT1	0.87	0.79–0.93	0.84	0.81	<0.001
PPARG	0.85	0.77–0.91	0.82	0.79	<0.001
SLC2A4	0.84	0.75–0.90	0.80	0.78	<0.001

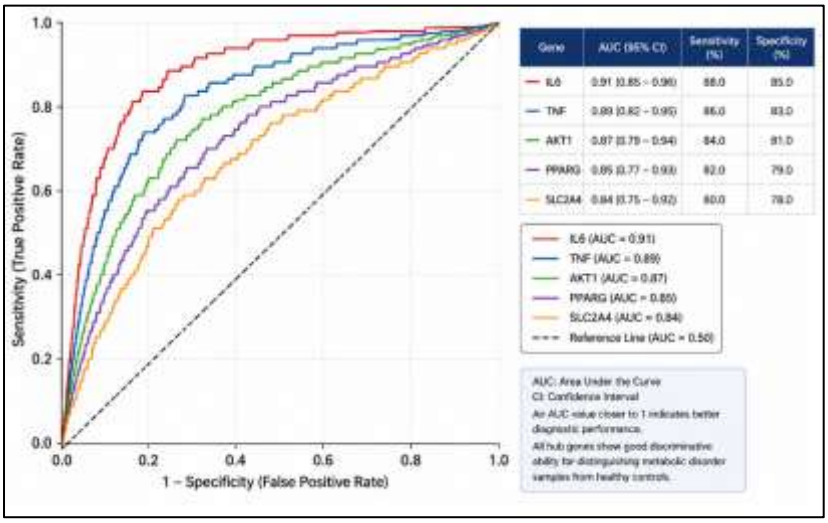

Fig. 5. Receiver operating characteristic (ROC) curve analysis of identified hub genes.

4.6 DISCUSSION

The current one utilized a transcriptomic and network-based method to determine the possible biomarkers of early diagnosis of metabolic disorders. The up-regulation of inflammatory genes like TNF and IL6 is observed and indicates the importance of chronic low-grade inflammation in metabolic dysfunction. This observation is in line with other reports that highlight inflammation as a major cause of insulin resistance and metabolic imbalance (Zhang, S., Zhou, J., Wu, W., Zhu, Y., and Liu, X.). (2023)). Equally, the PPARG and SLC2A4 down-regulation imply the disrupted glucose entry and lipid metabolism, which are characteristic of metabolic diseases. The fact that the PI3K–Akt and insulin signaling enrichment were also involved also speaks in favor of the role of these molecular mechanisms. The current study is more comprehensive in terms of identification of biomarkers in comparison with the existing literature, introducing a more extensive framework based on the combination of the differential expression analysis and the network and diagnostic validation. A major limitation of this study, however, is that the study makes use of a single publicly available dataset. The work should be corrected with validation of independent cohorts and experimental interventions (e.g., qPCR or Western blotting) that ensures clinical relevance of the identified biomarkers in future.

CONCLUSION

This paper has introduced a systemic transcriptomic and network model of identifying possible biomarkers that relate to metabolic disorders. The combination of the study of the differential gene expression, functional enrichment, construction of protein-protein interactions network, and diagnostic validation helped to recognize a list of essential genes, such as IL6, TNF, AKT1, PPARG, and SLC2A4 as the key factors in the development of the metabolic dysregulation. The patterns of expression and results of pathway enrichment observed emphasize the significant roles

of inflammatory signaling and the impaired pathways involving insulin in the disease development. The ROC analysis also indicated a very good diagnostic performance of these genes indicating that they may be useful as early detection biomarkers. Multi-gene signatures have a better ability to predict and greater robustness when compared to traditional single-marker methods. The key contribution of this paper is that it systematically combines bioinformatics and network-based approaches to discover biologically relevant and diagnostically important biomarkers out of transcriptomic data. The results are however limited due to use of one publicly available dataset. Research in the future on the validation of these biomarkers in relation to independent cohorts and any experimental method, like quantitative PCR and protein level assays, should be conducted to validate its clinical usefulness. Comprehensively, the present study has a credible basis when it comes to formulating early diagnostic methods and specific therapeutic treatment to metabolic disorders.

REFERENCES

1. Bhattacharya, P., Kanagasooriyan, R., & Subramanian, M. (2022). Tackling inflammation in atherosclerosis: Are we there yet and what lies beyond? *Current Opinion in Pharmacology, 66*, 102283. https://doi.org/10.1016/j.coph.2022.102283
2. Gao, X., Ke, C., Liu, H., Liu, W., Li, K., Yu, B., & Sun, M. (2017). Large-scale metabolomic analysis reveals potential biomarkers for early stage coronary atherosclerosis. *Scientific Reports, 7*, 11817. https://doi.org/10.1038/s41598-017-12254-1
3. Iida, M., Harada, S., & Takebayashi, T. (2019). Application of metabolomics to epidemiological studies of atherosclerosis and cardiovascular disease. *Journal of Atherosclerosis and Thrombosis, 26*(9), 747–757. https://doi.org/10.5551/jat.RV17036
4. Li, T., & Chiang, J. Y. L. (2009). Regulation of bile acid and cholesterol metabolism by PPARs. *PPAR Research, 2009*, 501739. https://doi.org/10.1155/2009/501739
5. Liu, X., Gao, J., Chen, J., Wang, Z., Shi, Q., Man, H., Guo, S., Wang, Y., Li, Z., & Wang, W. (2016). Identification of metabolic biomarkers in patients with type 2 diabetic coronary heart diseases based on metabolomic approach. *Scientific Reports, 6*, 30785. https://doi.org/10.1038/srep30785
6. Pathak, P., Liu, H., Boehme, S., Xie, C., Krausz, K. W., Gonzalez, F., & Chiang, J. Y. L. (2017). Farnesoid X receptor induces Takeda G-protein receptor 5 cross-talk to regulate bile acid synthesis and hepatic metabolism. *Journal of Biological Chemistry, 292*(26), 11055–11069. https://doi.org/10.1074/jbc.M117.784322
7. Poznyak, A., Grechko, A. V., Poggio, P., Myasoedova, V. A., Alfieri, V., & Orekhov, A. N. (2020). The diabetes mellitus–atherosclerosis connection: The role of lipid and glucose metabolism and chronic inflammation. *International Journal of Molecular Sciences, 21*(5), 1835. https://doi.org/10.3390/ijms21051835
8. Staels, B., & Fonseca, V. A. (2009). Bile acids and metabolic regulation: Mechanisms and clinical responses to bile acid sequestration. *Diabetes Care, 32*(Suppl 2), S237–S245. https://doi.org/10.2337/dc09-S355
9. Zhang, S., Zhou, J., Wu, W., Zhu, Y., & Liu, X. (2023). The role of bile acids in cardiovascular diseases: From mechanisms to clinical implications. *Aging and Disease, 14*(1), 261–282. https://doi.org/10.14336/AD.2022.0817