**Title: Identification of critical genes differentiating stable and unstable atherosclerotic plaques: a bioinformatic and computational analysis**

**Fellow:** Amirhossein Sahebkar, Biotechnology Research Center, Pharmaceutical Technology Institute, Mashhad University of Medical Sciences, Mashhad, Iran.

**Supervisor:** Raul D. Santos, Lipid Clinic Heart Institute (Incor), University of São Paulo, Medical School Hospital, São Paulo, Brazil.

**Abstract**

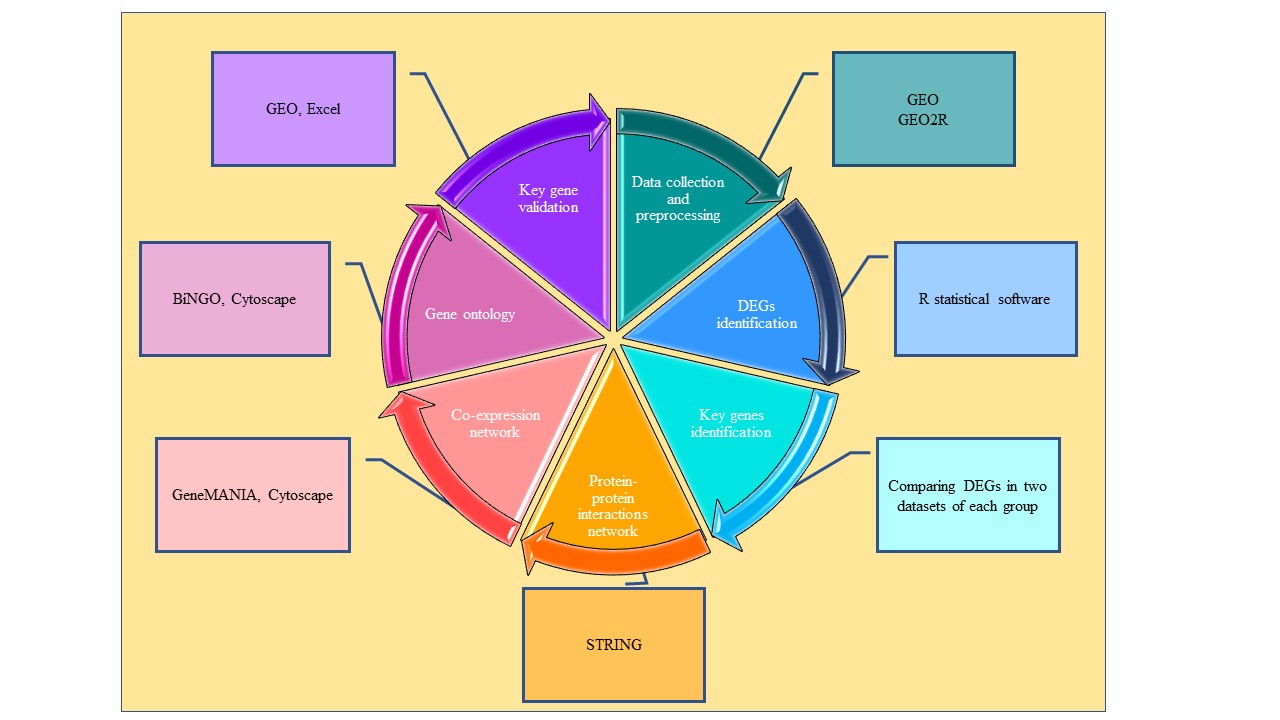
The rupture of unstable atherosclerotic plaques is critical step for acute events. Identification of blood biomarkers to distinguish between stable and unstable plaque formation would be a great achievement in this field. Differential expressed key genes related to formation of plaques could be an important approach to predict plaque vulnerability status and preventing the ensuing clinical events. In this study, we downloaded microarray profiles including GSE71226 and GSE20680 (group A: containing healthy vs. stable plaque samples) and GSE62646 and GSE34822 (group B: containing stable vs. unstable plaque samples) from GEO database which contains gene expression of blood samples. Differentially expressed genes were compared in both data sets of each group. We identified the key genes which have differential expression in the two datasets. Ten and 12 key genes were screened out in group A and B, respectively. GO enrichment was applied by the plugin “BiNGO” of the Cytoscape. The key genes were mostly enriched in biological process of positive regulation of cellular process. The protein-protein intraction and co-expression network was analyzed by the STRING and GeneMANIA plugin of Cytoscape, respectively, which showed that EGF, HBEGF, and MMP9 were at the core of the network. Further validation of key genes was conducted using two dataset showed that PDE5A and PROS1 were decreased in unstable plaques, while SOCS3, HBEGF and LILRB4 were increased. Overall, in our study, we used the several datasets to identify the key genes for the stable and unstable atherosclerotic plaque formation which have great potential as blood biomarkers.

**Key words:** Atherosclerosis, Stable plaque, Unstable plaque, Blood biomarkers

**Introduction**

Atherosclerosis is the most common cause of cardiovascular diseases (CVD) which are a growing cause of disability and death worldwide. Arterial wall inflammation plays a central role in atherosclerosis development [[1](#_ENREF_1)] [[2](#_ENREF_2)] [[3](#_ENREF_3)]. During the process fatty materials accumulate in arteries forming atherosclerotic plaques in intimal layers which eventually progress, thicken and harden the blood vessels. Immune cells accumulate within complicated lesions, propagate the disease and produce its complications [[4](#_ENREF_4)]. During this process plaque may grow preserving or not vessel lumen by the remodeling process. Over time, atherosclerosis may cause clinical manifestations through luminal narrowing and consequent blood flow restriction (stable angina in the heart) or by precipitating thrombi that acutely obstruct blood flow to the heart (acute coronary syndrome), brain (acute ischemic stroke), or lower extremities ( acute peripheral vascular disease episode) [[5](#_ENREF_5), [6](#_ENREF_6)]. As atherosclerosis is a predominantly asymptomatic condition, it is hard to find its presence especially on its early stages [[7](#_ENREF_7)] even in the presence of classical atherosclerosis predisposing factors. Bioinformatics-based approaches, particularly differential gene expression analysis, are an effective mean of diagnostic tools to identify potential molecular biomarkers for diagnosis in each stage of atherosclerosis.

We aimed to characterize differential blood gene expression from patients in groups with stable atherosclerotic plaques comparing with unstable plaques as well as in groups with stable atherosclerotic plaques comparing with healthy individuals. Therefore we characterized new blood biomarkers which could be applied to distinguish individuals in different stages of atherosclerosis with the aim of preventing subsequent events such as myocardial infraction and death. The workflow of the whole study is shown in Fig. 1.



**Figure 1.** Study workflow. The steps of this study include data collection and preprocessing, DEGs identification, Key genes identification, Protein-protein interactions network, Co-expression network, Gene ontology, and Key gene validation

**Methods**

*Data collection and preprocessing*

To explore the gene expression differences in plaque progression, four gene expression profiles were used. Gene expression profiles or genomic Spatial Event databases (GSEs) including GSE62646, GSE34822, GSE71226, and GSE20680 (Table 1) were downloaded from the Gene Expression Omnibus (Gene Expression Omnibus (GEO), <https://www.ncbi.nlm.nih.gov/geo>) and analyzed with GEO2R. The groups include patients with stable plaques, patients with unstable plaques and healthy individuals. The preprocessing of the series matrix profiles was conducted on R statistical software with related packages from Bioconductor (www.bioconductor.org).

**Table 1.** Gene expression profiles used in this study. The groups include patients with stable plaques, patients with unstable plaques and healthy individuals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Group** | **GEO accession** | **Type of sample** | **Groups/number of samples** |
| A | GSE62646 | Blood | Stable: 14  Unstable: 28 |
| GSE34822 | Blood | Stable: 16  Unstable: 16 |
| B | GSE71226 | Blood | Healthy: 3  Stable: 3 |
| GSE20680 | Blood | Healthy: 52  Stable: 141 |
| Validation | GSE41571 | Plaque | Stable: 6  Unstable: 6 |
| Validation | GSE120521 | Plaque | Stable: 3  Unstable: 3 |

*DEGs and Key gene identification*

For each GEO accession, differentially expressed genes (DEGs) between two groups were constructed. The series matrix files were annotated with official gene symbols using the platform files and annotation packages in the R software. The “LIMMA” package in the R was used to identify the DEGs. Genes with a │logFC│ > 0.5 (for GSE71226 and GSE20680 │logFC│ > 0.3 was applied) and adjusted p-value < 0.05 were considered to be DEGs. Common genes were identified between the DEGs of GSE71226 and GSE20680 identified as stable/healthy biomarker (group A) and the DEGs of GSE62646 and GSE34822 identified as stable/unstable biomarker (group B). The heatmap of these key genes was conducted using “ggplot2” package in R software. All the packages used in R software were downloaded from Bioconductor (http://www.bioconductor.org). Other related plots were created using GEO2R analysis. Moreover, function of key genes retrieved from GEO, were analyzed for each group.

*Interaction networks*

The STRING online tool (https://string-db.org/) was used for identification of protein-protein interactions of identified key genes. The STRING database covers the number of 67.6 million proteins from 14,094 organisms. It provides direct (physical) interactions and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases.

The interaction networks at the gene level were built by the GeneMANIA Cytoscape plugin. Gene co-expression network of key genes were constructed and Physical, co-expression, and pathway gene-gene interactions were evaluated.

*Gene ontology and pathway enrichment analysis.*

The plugin “BiNGO” tool in the “Apps” of the Cytoscape was used for the GO enrichment. Te q-value. The GO enriched the genes into 3 terms: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC).

*Key gene validation*

To further validation, GSE41571 and GSE120521were downloaded from GEO, and the key genes expression were analyzed and the key genes expression in the plaque dataset expression profiles was analyzed.

**Results**

*Identification of the DEGs*

The matrix file from GSE29111, GSE62646, and GSE34822 were pre-processed and annotated with the official gene symbol. Plots including volcano plot, gene expression value distribution for dataset, Q-Q plot, mean variance trend, and expression density curve were applied for data sets (Fig. 2).

The DEGs were identified using “LIMMA” package in the R software. The genes with p-value < 0.05 and │logFC│ > 0.5 (GSE62646, GSE34822, and GSE41571) or │logFC│ > 0.3 (GSE71226 and GSE20680) were considered DEGs. Eventually, DEGs of GSE29111, GSE62646, and GSE34822 were identified (Table 2).

**Table 2**. Number of identified DEGs in each data series.

|  |  |  |  |
| --- | --- | --- | --- |
| GEO accession | Number of identified DEGs | | |
| Upregulated genes | Downregulated genes | Total |
| GSE62646 | 209 | 179 | 388 |
| GSE34822 | 656 | 212 | 868 |
| GSE41571 | 1928 | 2109 | 4037 |
| GSE71226 | 1860 | 2589 | 4449 |
| GSE20680 | 48 | 30 | 78 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **E** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Figure 2.** Identification of DEGs in GSE62646 (1st row), GSE34822 (2nd row), GSE71226 (3rd row), and GSE20680 (4th row). A. volcano plot, B. Gene expression value distribution for dataset (Each box plot represents gene expression value of one patient sample.), C: Q-Q plot, D: mean variance trend, E: expression density curve.

*Identification of key genes*

Identification of key genes was conducted on the two groups to establish the common genes between the two datasets in each group. To figure out the key genes related to the patient with stable/unstable plaque,s we compared the 4449 DEG genes in GSE71226 with the 868 DEGs in GSE34822 and 12 key genes were figured out. To figure out the key genes related to the healthy cases/patient with stable plaque, we compared the 388 DEG genes in GSE62646 to the 78 DEGs in GSE20680 and 10 key genes were figured out (Table 3). The heat maps of these genes are shown in Fig. 3.

|  |  |
| --- | --- |
| A | B |

**Figure 3.** Heatmap of key genes. A: Heatmap of key genes in stable/unstable groups. B: Heatmap of key genes in healthy/stable groups.

**Table 3**. The list of Key genes.

|  |  |  |
| --- | --- | --- |
| **Gene name** | **Full name** | **Groups** |
| CBS | Cystathionine beta-synthase | A |
| CLC | Charcot-Leyden crystal galectin | A |
| CRYBB2 | Crystallin beta B2 | A |
| GPER1 | G protein-coupled estrogen receptor 1 | A |
| IGHM | Immunoglobulin heavy constant mu | A |
| MME | Membrane metalloendopeptidase | A |
| MMP9 | Matrix metallopeptidase 9 | A |
| PSPH | Phosphoserine phosphatase | A |
| EGF | Epidermal growth factor | B |
| ELOVL7 | ELOVL fatty acid elongase 7 | B |
| GUCY1A3 | Guanylate cyclase alpha-3 | B |
| GZMK | Granzyme K | B |
| HBEGF | Heparin binding EGF like growth factor | B |
| IL5RA | Interleukin 5 receptor subunit alpha | B |
| LILRB4 | Leukocyte immunoglobulin like receptor B4 | B |
| PDE5A | Phosphodiesterase 5A(PDE5A) | B |
| PF4 | Platelet factor 4 | B |
| PROS1 | Protein S | B |
| SH3PXD2B | SH3 and PX domains 2B | B |
| SOCS3 | Suppressor of cytokine signaling 3 | B |

*Gene Oncology (GO) enrichment of key genes*

GO enrichment was applied by the plugin “BiNGO” tool in the “Apps” of the Cytoscape. Te q-value. As shown in Fig. 4, these key genes were mostly enriched in Biological Process (BP) of positive regulation of cellular process (Fig. 4A); Cellular Components (CC) of extracellular region (Fig. 4B); and Molecular Functions (MF) of epidermal growth factor receptor binding (Fig. 4D).

|  |
| --- |
| A |
| B |
| C |

**Figure 4.** GO enrichment of key genes. A: Biological Process of GO enrichment. B. Cellular components of GO enrichment. C. Molecular function of GO enrichment;

*Interaction networks of the key genes*

The PPI was analyzed by the STRING online tool (website: https://string-db.org/). Overall, 22 key genes were analyzed. As a result, the 22 genes formed a network with 20 nodes and 13 edges with PPI enrichment p-value of 2.95e-06 (Fig. 5). EGF, HBEGF, and MMP9 were at the core of the network. A co-expression network for the key genes was constructed, with gene-correlation interactions consisting of 39 nodes and 143 edges, by mapping the key genes onto a massive database of functional-interaction datasets in the GeneMANIA plugin of Cytoscape (Fig. 6).

|  |
| --- |
|  |

**Figure 5.** PPI network for the 72 hub genes. The circles represent the proteins encoded by the corresponding genes; lines represent the interactions between the proteins.

|  |  |
| --- | --- |
|  |  |

**Figure 6.** Gene-gene interaction network for 22 key genes. A set of key genes were provided as a query (black nodes) and a number of additional genes were predicted to be related (gray nodes).

*Validation of key genes in* *atherosclerotic plaques*

Because of importance of unstable plaque formation, to further validate the key genes in identification of stable/unstable cases (group B), the human plaques expression profile of dataset GSE120521 and GSE41571 was downloaded from GEO, and the key genes expression were analyzed. We detected the expression of the 12 key genes in both profiles. Expression profiling of GSE41571 (6 stable and 5 unstable) was done by microarray profile. In GSE41571, GUCY1A3, PDE5A, and PROS1 were decreased in unstable plaques, while SOCS3, LILRB4, PF4, HBEGF, and ELOVL7 were increased (Fig. 7-A). Expression profiling of GSE120521 (4 stable and 4 unstable) was done by high throughput sequencing. In GSE120521, PDE5A and PROS1 were decreased in unstable plaques, while SOCS3, HBEGF and LILRB4 were increased (Fig.7-B).

|  |  |
| --- | --- |
|  |  |

**Figure 7.** The expression of 12 key genes of group B in atherosclerotic plaque. A. The expression in GSE41571 dataset. B. The expression in GSE120521 dataset.

**Discussion**

In this study, we aimed to identified key genes which were differentially expressed in blood in different atherosclerosis stages. We compared expressed genes in two groups; group A: healthy controls vs. patients with stable plaque and group B: patients with stable plaque vs. patients with unstable plaques. We used four microarray datasets including GSE62646 and GSE34822 (in group A) and GSE71226 and GSE20680 (in group B) to identified key genes which were differentially expressed in both datasets. Ten and 12 key genes were screened out from these datasets in groups A and B, respectively. Further GO enrichment of the 22 key genes indicates these genes to be mostly related to regulation of cellular process, serine family amino acid metabolic process, epidermal growth factor receptor signaling pathway and JAK-STAT cascade. These processes are related to the key pathogenic steps of atherosclerosis progression. PPI enrichment showed that EGF, HBEGF, and MMP9 were at the core of the network. Association of these genes with atherosclerosis has been found. EGF is able to modulate lipoprotein fractions in human fetal intestine and hepatic-derived cell lines. Berrahmoune et al. showed that plasma EGF concentrations were significantly and positively correlated with apolipoprotein A1 and HDL-cholesterol concentrations. In addition, EGF concentrations in PBMCs were negatively correlated with apolipoprotein B, total and low-density lipoprotein-cholesterol, and triglyceride concentrations [[8](#_ENREF_8)]. HBEGF has been suggested to induce the migration and proliferation of vascular smooth muscle cells. Matsumoto et al. found that plasma HB-EGF levels increase in parallel with fat accumulation in human, and (4) the subjects with coronary artery disease have higher plasma HB-EGF levels, associated with fat accumulation. Moreover, it was showed that HB-EGF antisense also induced an effective suppression of atherosclerosis development [[9](#_ENREF_9), [10](#_ENREF_10)]. Synthesis of MMPs has been reported in coronary atherosclerotic lesions in patients with unstable angina. Kai et al. demonstrated that plasma MMP-9 was transiently elevated in patients with unstable angina and acute myocardial infarction [[11](#_ENREF_11)]. Moreover, found MMP9 in blood as markers for early atherosclerosis [[12](#_ENREF_12)].

Furthermore, we investigated the expression of identified key genes in atherosclerotic plaque samples of GSE41571 and GSE120521. It was found that the expression of PDE5A and PROS1 were decreased, while SOCS3, HBEGF and LILRB4 were increased in unstable plaques. So, these genes may have critical roles in plaque formation. Hurtado et al. observed decrease in PROS1 mRNA expression in atherosclerotic carotids compared to the normal ones [[13](#_ENREF_13)]. Yang et al. showed that inhibition of JAK2/STAT3/SOCS3 signaling attenuates atherosclerosis [[14](#_ENREF_14)]. Nakata et al. demonstrated that in atherosclerotic plaques of coronary arteries with eccentric intimal thickening, both SMCs and macrophages in and around the core lesions, in addition to the intimal and medial SMCs located adjacent to the plaque, produced HB-EGF protein [[15](#_ENREF_15)]. Jiang et al. showed LILRB4 deficiency aggravates the development of atherosclerosis and plaque instability by increasing the macrophage inflammatory response via NF-κB signaling [[16](#_ENREF_16)]. It was also shown that LILRB4 deficiency, significantly accelerated development of atherosclerotic lesions and increased instability of plaques associated with increased infiltration of lipids and decreased collagen components and smooth muscle cells were observed [[17](#_ENREF_17)]. In spite of SOCS3 and HBEGF, LILRB4 showed negative correlation with plaque formation. Theses might because of the protection role of LILRB4 against plaque formation, which caused the increase of LILRB4 in order to inhibit atherosclerotic plaque progression.

Overall, in our study, we used the several datasets to identify the key genes for the stable and unstable atherosclerotic plaque formation which have great potential as blood biomarkers. Key genes may play crucial roles in inflammation, fibrous cap formation and degradation, increased infiltration of lipids and decreased collagen components and smooth muscle cells, increased instability of plaques and lead to the rupture of atherosclerotic plaque.

**References**

1. Bertrand MJ, Tardif JC. Inflammation and beyond: new directions and emerging drugs for treating atherosclerosis. Expert Opin Emerg Drugs. 2017;22(1):1-26.

2. Cardiovascular disease. Nat Biotechnol. 2000;18 Suppl:IT15-7.

3. Kypreos KE, Bitzur R, Karavia EA, Xepapadaki E, Panayiotakopoulos G, Constantinou C. Pharmacological Management of Dyslipidemia in Atherosclerosis: Limitations, Challenges, and New Therapeutic Opportunities. Angiology. 2018.

4. Soehnlein O, Libby P. Targeting inflammation in atherosclerosis — from experimental insights to the clinic. Nature Reviews Drug Discovery. 2021;20(8):589-610.

5. Poznyak AV, Nikiforov NG, Markin AM, Kashirskikh DA, Myasoedova VA, Gerasimova EV, et al. Overview of OxLDL and Its Impact on Cardiovascular Health: Focus on Atherosclerosis. Frontiers in Pharmacology. 2021;11(2248).

6. Bentzon JF, Otsuka F, Virmani R, Falk E. Mechanisms of Plaque Formation and Rupture. Circulation Research. 2014;114(12):1852-66.

7. Fowkes FGR, Aboyans V, Fowkes FJ, McDermott MM, Sampson UK, Criqui MH. Peripheral artery disease: epidemiology and global perspectives. Nature Reviews Cardiology. 2017;14(3):156.

8. Berrahmoune H, Lamont JV, Herbeth B, Lambert D, Masson C, McPhillips M, et al. Association between EGF and lipid concentrations: A benefit role in the atherosclerotic process? Clinica Chimica Acta. 2009;402(1):196-8.

9. Kim S, Graham MJ, Lee RG, Yang L, Kim S, Subramanian V, et al. Heparin-binding EGF-like growth factor (HB-EGF) antisense oligonucleotide protected against hyperlipidemia-associated atherosclerosis. Nutrition, Metabolism and Cardiovascular Diseases. 2019;29(3):306-15.

10. Lee S, Lee R, Graham M, Kim S-G, Yang L, Rateri D, et al. Targeting of Heparin Binding EGF-like Growth Factor (HB-EGF) Efficiently Inhibits Atherosclerosis. Arteriosclerosis, Thrombosis, and Vascular Biology. 2015;35(suppl\_1):A358-A.

11. Kai H, Ikeda H, Yasukawa H, Kai M, Seki Y, Kuwahara F, et al. Peripheral blood levels of matrix metalloproteases-2 and -9 are elevated in patients with acute coronary syndromes. Journal of the American College of Cardiology. 1998;32(2):368-72.

12. Söder PÖ, Meurman JH, Jogestrand T, Nowak J, Söder B. Matrix metalloproteinase-9 and tissue inhibitor of matrix metalloproteinase-1 in blood as markers for early atherosclerosis in subjects with chronic periodontitis. Journal of Periodontal Research. 2009;44(4):452-8.

13. Hurtado B, Muñoz X, Recarte-Pelz P, Garcia N, Luque A, Krupinski J, et al. Expression of the vitamin K-dependent proteins GAS6 and protein S and the TAM receptor tyrosine kinases in human atherosclerotic carotid plaques. Thrombosis and haemostasis. 2011;105(05):873-82.

14. Yang X, Jia J, Yu Z, Duanmu Z, He H, Chen S, et al. Inhibition of JAK2/STAT3/SOCS3 signaling attenuates atherosclerosis in rabbit. BMC Cardiovasc Disord. 2020;20(1):133.

15. Nakata A, Miyagawa J-i, Yamashita S, Nishida M, Tamura R, Yamamori K, et al. Localization of Heparin-Binding Epidermal Growth Factor–Like Growth Factor in Human Coronary Arteries. Circulation. 1996;94(11):2778-86.

16. Jiang Z, Qin J-J, Zhang Y, Cheng W-L, Ji Y-X, Gong F-H, et al. LILRB4 deficiency aggravates the development of atherosclerosis and plaque instability by increasing the macrophage inflammatory response via NF-κB signaling. Clinical Science. 2017;131(17):2275-88.

17. Kretzschmar F, Piecha R, Jahn J, Potru PS, Spittau B. Characterization of the Leucocyte Immunoglobulin-like Receptor B4 (Lilrb4) Expression in Microglia. Biology. 2021;10(12):1300.