Atherosclerosis 291 (2019) 34–43

Contents lists available at ScienceDirect

Atherosclerosis

journal homepage: www.elsevier.com/locate/atherosclerosis

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

Xue-bin Wang[a], Ning-hua Cui[b], Xia'nan Liu[a], Liang Ming[a,*]

[a] Department of Clinical Laboratory, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
[b] Zhengzhou Key Laboratory of Children's Infection and Immunity, Children's Hospital Affiliated to Zhengzhou University, Zhengzhou, China

HIGHLIGHTS

- We derived a 12-gene expression score (GES12) by systematic use of gene network construction and Support Vector Machine (SVM) algorithm.
- GES12 showed a reliable accuracy for predicting a broad spectrum of coronary stenosis in multiple cohorts.
- A GES12-based nomogram may improve the performance of GES12 in predicting coronary stenosis.

ARTICLE INFO

Keywords:
Gene expression score
Coronary artery stenosis
Gene coexpression network
Prediction model

ABSTRACT

Background and aims: We aimed to identify a blood-based gene expression score (GES) to predict the severity of coronary artery stenosis in patients with known or suspected coronary artery disease (CAD) by integrative use of gene network construction, Support Vector Machine (SVM) algorithm, and multi-cohort validation.
Methods: In the discovery phase, a public blood-based microarray dataset of 110 patients with known CAD was analyzed by weighted gene coexpression network analysis and protein-protein interaction network analysis to identify candidate hub genes. In the training set with 151 CAD patients, bioinformatically identified hub genes were experimentally verified by real-time polymerase chain reaction, and statistically filtered with the SVM algorithm to develop a GES. Internal and external validation of GES was performed in patients with suspected CAD from two validation cohorts (n = 209 and 206).
Results: The discovery phase screened 15 network-centric hub genes significantly correlated with the Duke CAD Severity Index. In the training cohort, 12 of 15 hub genes were filtered to construct a blood-based GES12, which showed good discrimination for higher modified Gensini scores (AUC: 0.798 and 0.812), higher Sullivan Extent scores (AUC: 0.776 and 0.778), and the presence of obstructive CAD (AUC: 0.834 and 0.792) in two validation cohorts. A nomogram comprising GES12, smoking status, hypertension status, low density lipoprotein cholesterol level, and body mass index further improved performance, with respect to discrimination, risk classification, and clinical utility, for prediction of coronary stenosis severity.
Conclusions: GES12 is useful in predicting the severity of coronary artery stenosis in patients with known or suspected CAD.

1. Introduction

Coronary artery disease (CAD) is the leading cause of death worldwide [1]. In symptomatic patients, assessment of the severity of coronary artery stenosis is critical for describing coronary anatomy, diagnosing obstructive CAD, and deciding the indication for myocardial revascularization [2]. Cardiac imaging methods, such as invasive coronary angiography and noninvasive multidetector CT angiography, reveal the location and extent of coronary obstructive lesions, which are strong predictors for long-term outcomes [2,3]. However, the need for specialized centers and the risk for radiation exposure limit their routine use in large-scale population screenings [4,5]. Blood-based

* Corresponding author. Department of Clinical Laboratory, The First Affiliated Hospital of Zhengzhou University, Henan, China.
E-mail address: mingliangjyk2011@163.com (L. Ming).

https://doi.org/10.1016/j.atherosclerosis.2019.10.001
Received 1 August 2019; Received in revised form 25 September 2019; Accepted 8 October 2019
Available online 09 October 2019
0021-9150/

biomarkers, including C-reactive protein and other inflammatory markers, received increasing attention in this regard but only exhibited modest effects in gaining predictive performance [6,7], highlighting the necessity of identifying novel molecular signatures to reflect the severity of coronary artery stenosis [8,9].

Recently, developing a blood-based gene expression score (GES) has been considered as a promising biomarker strategy for prediction of coronary atherosclerosis [10]. However, despite great efforts in this field, constructing a reproducible gene signature has proved difficult, due to lack of validation in independent cohorts, batch variability in multiple experimental platforms, and low levels of differential expression of single genes [11]. Of note, established GESs, including a 9-gene signature identified by Muse et al. [12], a 23-gene classifier (i.e. CorusCAD) reported by Rosenberg et al. [13], and 238 gene profiles reported by Kim et al. [14], mainly served for discrimination of CAD patients from non-CAD controls, but few were initially designed for reflection of coronary stenosis severity [15]. Moreover, previous studies established GESs typically by screening the differentially expressed genes (DEGs) individually associated with CAD, but barely investigated whether the panel of functionally co-expressed genes had the potential to predict the severity of coronary artery stenosis [16].

Hence, by performing the weighted gene coexpression network analysis (WGCNA) in a whole-blood gene expression profile retrieved from the Gene Expression Omnibus (GEO) database, we constructed gene modules within network and screened 15 network-centric genes associated with the Duke CAD Severity Index. Then, after verification by real-time polymerase chain reaction (RT-PCR) in a training set of 151 CAD patients, 12 of the 15 genes were selected to generate a blood-based 12-gene expression score (GES12), which showed good calibration and discriminative ability to predict coronary stenosis severity in two validation cohorts. Finally, a nomogram, composed of GES12 and four traditional cardiovascular risk factors, was built to further improve risk reclassification and clinical usefulness of GES12 in predicting coronary stenosis severity.

2. Materials and methods

2.1. The discovery set searching from the GEO database

We searched the gene expression profiles from the GEO database by using the combinations of the following keywords: ("coronary artery stenosis" or "CAD" or "myocardial infarction") and ("whole blood" or "peripheral blood") and "expression profiling" and "*Homo sapiens*". After excluding irrelevant and small-sized datasets ($n < 50$), the GSE12288, as the only dataset that deposited the data of coronary stenosis severity (i.e. Duke CAD Severity Index) in 110 CAD patients and 110 controls [17], was selected as the discovery set. Considering that the Duke Index was mainly designed for scoring the severity of coronary stenosis in patients with known CAD, but not in subjects with luminal stenosis of < 50% [18,19], we only integrate the raw data of 110 CAD patients for WGCNA modeling. The whole blood gene expression data of GSE12288 were first analyzed using the Affymetrix Human Genome U133A platform, and then normalized with RMA method (R package *affy*) [20]. When multiple probe sets corresponded to the same gene symbol, the median intensity across the probe sets was calculated as the expression level of the corresponding gene, which resulted in a total of 13,515 genes being annotated.

2.2. WGCNA construction

Out of 13,515 genes, 8618 were further selected for WGCNA modeling based on the following criteria: 1) DEGs associated with the Duke Index at a false discovery rate (FDR) of < 0.05; 2) genes with high expression variance (top 50%) [21]. In WGCNA, the pairwise correlations among the 8618 genes were first calculated to build a matrix of adjacencies, based on a power function that quantified the degree of correlations between two genes. Then, the β value, as a weighted coefficient emphasizing stronger pairwise correlations and penalizing weaker ones, was estimated for weighting the adjacency matrix [22]. Finally, to generate modules with highly regulated genes, we performed average linkage hierarchical clustering to group 8618 analyzed genes according to the topological overlap of gene interconnectivity, followed by using a dynamic tree cut algorithm to merge the branches of the network dendrogram into gene modules [23]. To ensure the stability of the clustering process, we removed outlier samples which were radically different from other samples in gene expression patterns [24]. Each module had a minimum size of 30 genes [22], and was assigned for different colors for visualization in the dendrogram.

2.3. Clinical significance and functional annotations of identified models

For each module, the principal component analysis was performed to choose the primary component of a given module, termed the eigengene, which was the representative explaining the largest proportion of module variance [22]. To assess module-trait correlations, multivariable linear regression was conducted with the module eigengene as the independent variable and the Duke Index as the predictor of interest (adjusting for age and sex). Gene modules significantly related to the Duke Index (i.e. differential module [DM]) were uploaded into the Comparative Toxicogenomics Database to summarize the enrichment of functional annotations and gene-disease interactions [25]. In WGCNA, the selection of hub genes in the DM was based on a series of predefined thresholds: module membership > 0.75, |correlation coefficient| with the Duke Index > 0.2, FDR < 0.5. After mapping gene contents into the STRING database [26], we constructed the protein-protein interaction (PPI) network of the DM, which screened genes with connectivity degree of > 4 as the hub genes [24].

2.4. Training set for developing a GES using the bioinformatically identified hub genes

A training set, including 151 patients with known CAD, was enrolled from the First Affiliated Hospital of Zhengzhou University (ZZUFH) at Henan province, northern China between August 2017 and December 2017 (full details shown in Supplementary Materials and Methods and Supplementary Table 1). CAD was defined as luminal stenosis of ≥50% in ≥1 major coronary artery by invasive angiography, with typical symptoms of angina, typical electrocardiographic patterns, or/and negative alterations in cardiac biomarkers. For ensuring the comparability of the results between discovery set and training set, we still used the Duke Index to assess the severity of coronary stenosis for each patient. We used a Support Vector Machine (SVM) algorithm to select the essential contents of a GES from the bioinformatically identified hub genes. The GES was computed by weighting the expression level of a given gene with the corresponding regression coefficient and then taking the sum across all included genes [27]. Supplementary Table 2 presents a glossary of key technical terms related to the construction of GES.

2.5. Internal and external validation of GES12

Validation cohorts included two cross-sectional studies of consecutive individuals undergoing coronary angiography for suspected CAD: 1) internal validation set with 209 patients also recruited from ZZUFH between May 2018 and December 2018; 2) external validation set including 206 subjects enrolled from Zhongnan Hospital of Wuhan University (WHUZH) at Hubei province, central China between October 2016 and March 2017 (full details shown in Table 1 and Supplementary Materials and Methods) [28,29]. The suspected CAD was defined as: 1) patients with a history of chest pain; 2) patients with suspected symptoms of angina, including intermittent arrhythmias in the absence of a positive stress test or an equivocal stress test for myocardial ischemia;

Table 1
Clinical data of discovery, training, and validation cohorts.

Variables	Discovery set (GSE12288, n = 110)	Training set (ZZUFH, n = 151)	Internal validation (ZZUFH, n = 209)	External validation (WHUZH, n = 206)
Age, years	54.6 ± 7.1	57.6 ± 5.2	62.1 ± 10.7	63.6 ± 9.4
Male, n (%)	88 (80.0)	76 (50.3)	128 (61.2)	117 (56.8)
BMI, kg/m^2	-	24.9 ± 3.3	25.0 ± 3.4	24.9 ± 3.4
Smoking, n (%)	-	49 (32.5)	68 (32.5)	72 (35.0)
Drinking, n (%)	-	47 (31.1)	57 (27.3)	56 (27.2)
Hypertension, n (%)	-	62 (41.1)	109 (52.2)	104 (50.5)
Diabetes, n (%)	-	47 (31.1)	54 (25.8)	57 (27.7)
TC, mmol/L	-	5.25 ± 0.83	5.23 ± 0.90	5.25 ± 1.01
TG, mmol/L	-	1.55 ± 0.65	1.55 ± 0.88	1.45 ± 0.72
LDL-c, mmol/L	-	3.66 ± 0.84	3.57 ± 0.84	3.65 ± 0.80
HDL-c, mmol/L	-	1.11 ± 0.19	1.12 ± 0.19	1.08 ± 0.20
Duke Index [a]	42 (32–63)	42 (32–63)	-	
Modified Gensini score [a]	-	-	31 (21.5–51)	27 (17–50.5)
Sullivan score [a]	-	-	30.5 (21–49)	29 (17–45)
Obstructive CAD, n (%)	-	-	152 (72.7)	136 (66.0)
Revascularization event, (%)		-	123 (58.9)	120 (58.3)

BMI: body mass index; TC: total cholesterol; TG: triglyceride; LDL-c: low-density lipoprotein cholesterol; HDL-c: high-density lipoprotein cholesterol; CAD: coronary artery disease.
[a] Values were expressed as median (interquartile range) due to skewed distributions.

3) patients with high levels (≥3) of cardiovascular risk factors; 4) no known prior myocardial infarction, revascularization, or obstructive CAD [30]. The degree of luminal narrowing was quantified by modified Gensini score and Sullivan Extent score [31,32]. Each participant of two cohorts was followed for revascularization (percutaneous or surgical) in 30 days after angiography. Characterization of atherosclerotic plaques was assessed in a sub-study of the internal validation cohorts with 152 participants who underwent single-vessel gray-scale and virtual histology (VH) intravascular ultrasound (IVUS) after diagnostic angiography. The IVUS region of interest was the most diseased segment, defined as the greatest plaque volume in contiguous across sections over an axial distance of 10 mm [33]. The training and validation sets of the study were approved by the local ethics committees; participants signed written informed consents accordingly.

2.6. Statistical analysis

The area under the receiver-operator characteristics (ROC) curve (AUC) was calculated to assess the performance of GES12 in predicting the severity of coronary stenosis (as determined by the Duke Index, modified Gensini scores and Sullivan Extent scores) and the presence of obstructive CAD. For calculating the AUCs, variables on coronary angiographic scores were dichotomized using their median values as the cutoff points. The calibration performance of GES12 was assessed by calibration curves, followed by the Hosmer-Lemeshow test.

Effect modification by cardiovascular risk factors, including age, sex, smoking status, alcohol drinking status, BMI, blood lipids, hypertension, and diabetic status, was tested by a ROC curve regression analysis to incorporate interaction terms (GES12 × covariate) for modeling the effect of stratification of covariates on performance of GES12 [34]. The significance of GES12 and cardiovascular risk factors in the training set was evaluated by stepwise multivariate logistic regression for identifying the independent risk factors of higher Duke Index. The regression coefficients of identified variables were used to construct a nomogram, which was a 0- to 160-point scale with predicted probabilities. The incremental predictive value of the nomogram compared to the GES12 was assessed by the AUCs, net reclassification improvement (NRI), and integrated discrimination improvement (IDI) [35]. For calculating NRI and IDI, the risk estimates of nomogram and GES12 were categorized as 0–20%, 20–50%, and ≥50%, corresponding to low, intermediate, and high risk of having higher coronary angiographic scores, respectively [13]. The clinical utility of the nomogram and GES12 was compared using the decision curve analysis, which plotted the net benefit of two models at different threshold probabilities [36]. All analyses were conducted with R software (version 3.5.0). A two-sided p value of < 0.05 was considered significant.

3. Results

3.1. Network construction by WGCNA in the discovery set

A workflow of the study is shown in Supplementary Fig. 1. At first, the process of average linkage hierarchical clustering identified an outlier sample (GSM308690), which was removed from further analysis (Supplementary Fig. 2A). A total of 109 CAD samples, with gene expression profiles of 8618 genes, were finally included into the WGCNA approach. A β value of 3, as the first value that gave an R^2 of ≥0.9, was selected as the soft-thresholding to fulfill the scale-free approximation criterion (Supplementary Fig. 2B). In total, 11 coexpression modules were identified, with module sizes ranging from 42 genes to 2603 genes (Fig. 1A and Supplementary Fig. 3). For each module, the eigengene was regressed in multivariable linear regression, which showed that the brown module eigengene level was significantly associated with the Duke Index (R = 0.26, FDR = 0.007, Fig. 1B). Functional enrichment analyses of the brown module annotated seven overrepresented biological processes in gene ontology. Most annotations were CAD-associated, involving inflammation, oxidative stress, nitric oxide biosynthesis, etc (Fig. 1C). Gene-disease interaction analyses mapped module genes into several terms of cardiovascular disease, including “Atherosclerosis”, “Vascular Disease”, “Coronary Disease”, etc (Fig. 1D). Taken together, we selected the brow module (with 1148 gene contents) as the DM for hub gene selection.

3.2. Identification of hub genes from the brown module

Based on predefined thresholds, WGCNA and PPI networks identified 30 and 47 hub genes, respectively, in the brown module (Supplementary Table 3 and Supplementary Fig. 4). By selecting common hub genes between WGCNA and PPI networks, we further identified 15 genes, which simultaneously showed significant correlations with the Duke Index and high degree of interconnectivity (Table 2 and Supplementary Fig. 5A). In clustering analyses, the expression

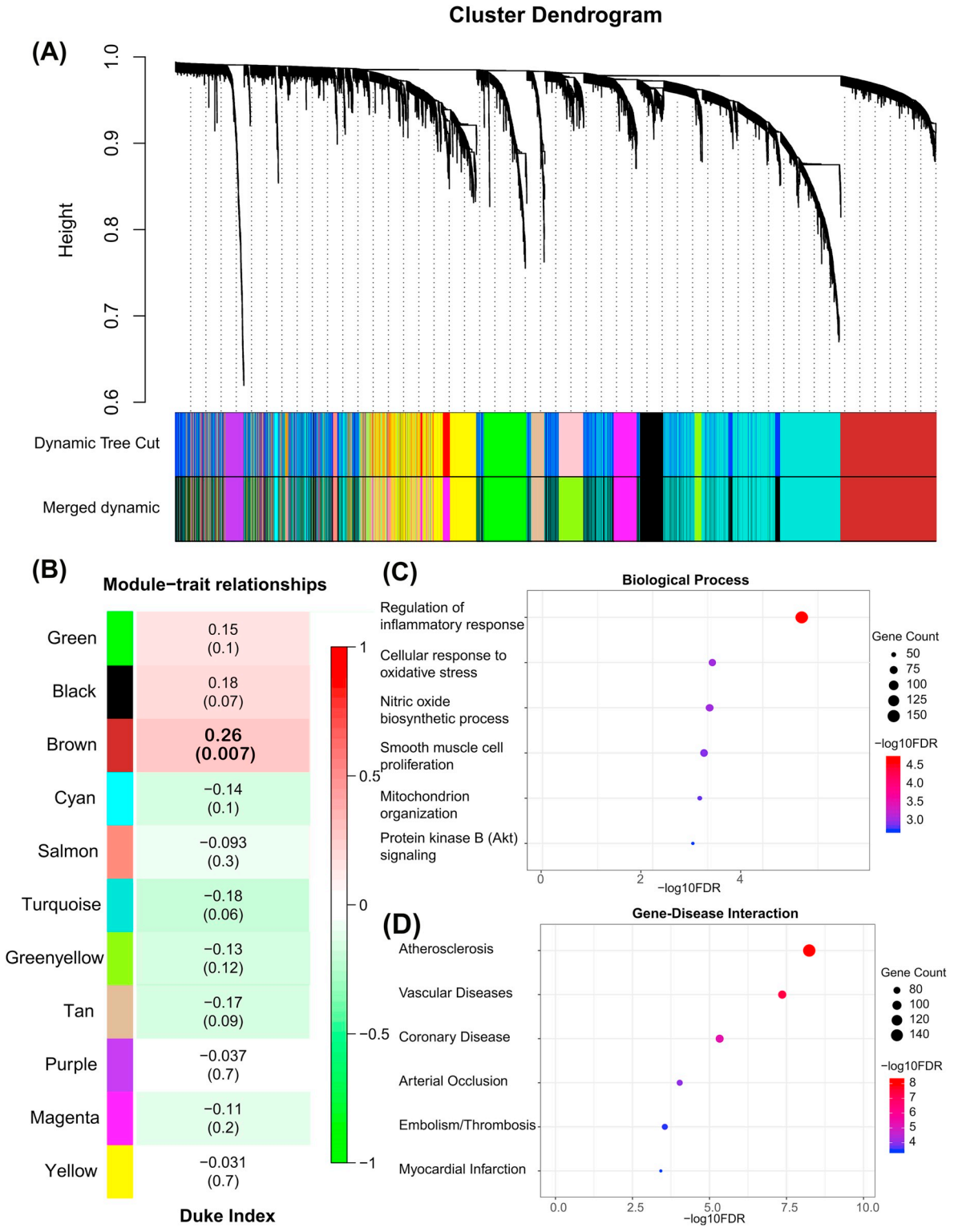

Fig. 1. Identification of a differential module associated with the Duke Index in the discovery phase.
(A) Network dendrogram and colors of identified modules based on a dissimilarity measure (1-TOM). (B) Correlations of module eigengenes and Duke Index. (C) Biological processes overrepresented by genes in the brown module. (D) Cardiovascular diseases enriched by genes in the brown module. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article).

Table 2
Associations of 15 bioinformatically identified hub genes with Duke Index in CAD patients from discovery and training cohorts.

Gene symbol	Chr. No.	Discovery cohort (GSE12288, n = 109)				Training cohort (Higher Duke Index *vs.* Lower Duke Index, n = 151)					
		Measures of interconnectivity		Correlations with Duke Index		Without adjustment			With adjustment [b]		
		Module membership [a]	Connectivity degree [a]	Spearman Coefficient	FDR	β	OR (95%CI)	FDR	β	OR (95%CI)	FDR
GUCY1A3	4	0.822	22	0.349	2.01E-04	0.748	2.11 (1.46–3.06)	**6.76E-04**	0.696	2.01 (1.27–3.16)	**0.015**
AAMP	2	0.821	21	0.342	2.76E-04	0.758	2.14 (1.46–3.12)	**6.76E-04**	0.747	2.11 (1.33–3.35)	**0.015**
P2RX7	12	0.871	42	0.388	3.13E-05	0.701	2.02 (1.40–2.91)	**8.29E-04**	0.554	1.74 (1.13–2.69)	**0.028**
PINK1	1	0.864	37	0.319	7.09E-04	0.590	1.80 (1.27–2.56)	**0.002**	0.580	1.79 (1.15–2.77)	**0.028**
TLL1	4	0.832	22	0.332	4.25E-04	0.592	1.81 (1.26–2.60)	**0.002**	0.502	1.65 (1.08–2.52)	**0.038**
XDH	2	0.822	14	0.383	3.88E-05	0.600	1.82 (1.27–2.62)	**0.002**	0.505	1.66 (1.06–2.58)	**0.043**
FOLH1	11	0.841	17	0.323	0.001	0.616	1.85 (1.26–2.71)	**0.004**	0.475	1.61 (1.04–2.48)	**0.043**
CCL18	17	0.755	14	0.329	4.83E-04	0.497	1.64 (1.16–2.33)	**0.008**	0.493	1.64 (1.11–2.42)	**0.028**
AIF1	6	0.806	13	0.364	1.00E-04	0.412	1.51 (1.07–2.12)	**0.025**	0.401	1.49 (1.03–2.16)	**0.043**
CNNM2	10	0.828	11	0.361	1.16E-04	0.383	1.47 (1.05–2.05)	**0.031**	0.404	1.50 (1.03–2.18)	**0.043**
PARP1	1	−0.890	61	−0.302	0.014	−0.709	0.49 (0.34–0.72)	**8.29E-04**	−0.463	0.46 (0.30–0.72)	**0.011**
KLF4	19	−0.834	25	−0.285	0.003	−0.537	0.58 (0.41–0.84)	**0.005**	−0.499	0.61 (0.41–0.90)	**0.028**
CNR2	4	0.813	26	0.320	6.93E-04	0.351	1.41 (1.01–1.99)	0.052	0.227	1.26 (0.84–1.88)	0.272
FANCC	9	0.827	20	0.382	1.76E-04	0.315	1.37 (0.98–1.91)	0.063	0.332	1.39 (0.94–2.07)	0.098
ALOXE3	17	0.756	12	0.342	2.69E-04	0.060	1.06 (0.77–1.47)	0.713	0.038	1.04 (0.68–1.60)	0.865

Bold values indicate FDR < 0.05.
[a] Module membership and connectivity degree are measures of interconnectivity derived from WGCNA and PPI networks, respectively.
[b] Adjusted for age, sex, BMI, smoking status, alcohol drinking status, TC, TG, LDL-c, HDL-c, hypertension, and diabetes.

pattern of the 15 genes clearly separated 109 CAD patients into two sample clusters (Supplementary Fig. 5B). A significant difference in the Duke Index between two clusters was identified (p = 2.21E-5, Supplementary Fig. 5C). We finally chose these 15 genes for further analysis.

3.3. Construction of a GES12 for predicting the severity of coronary stenosis in a training set

In a training set of 151 CAD patients from Henan province, northern China, we quantified the expression of 15 bioinformatically identified hub genes using RT-PCR. After dividing CAD patients into two groups based on the median value of the Duke Index (> 42 *vs.* ≤ 42), the SVM algorithm revealed that the combination of 12 genes provided the best performance in predicting higher Duke Index among all 2 to 15 gene combinations (n = 32,752) of the 15 analyzed genes (predictive accuracy = 78.2%, Kappa value = 0.452, Fig. 2A and Supplementary Tables 4 and 5). The expression levels of all 12 genes were independently associated with higher Duke Index in logistic regression with and without adjustment for confounders (FDR < 0.05, Table 2). Then, to test whether there was a risk of overfitting in this 12-gene signature, we performed a sensitivity analysis by sequential removal of each gene to construct all possible twelve-minus-one gene signatures [37]. ROC curves showed that none of the twelve-minus-one signatures had larger AUCs than the 12-gene signature, suggesting that each of the 12 genes was essential for predicting higher Duke Index (Fig. 2B). Finally, to control for the effect sizes of 12 genes, we built a risk score model (GES12) by weighting the expression level of 12 genes with corresponding coefficients derived from the univariate logistic regression as follows: $GES12 = 0.748 \times GUCY1A3 + 0.758 \times AAMP + 0.701 \times P2RX7 + 0.590 \times PINK1 + 0.592 \times TLL1 + 0.600 \times XDH + 0.616 \times FOLH1 + 0.497 \times CCL18 + 0.412 \times AIF1 + 0.383 \times CNNM2 - 0.709 \times PARP1 - 0.537 \times KLF1$. The AUC for the GES12 was 0.775 (95% confidence interval [CI]: 0.699–0.851) for prediction of higher Duke Index (Fig. 2C). The optimal cut point of the GES12 was 4.983, corresponding to a maximum Youden index of 0.460, a specificity of 0.747, and a sensitivity of 0.713 (Supplementary Table 6). When we classified patients with a risk score of > 4.983 into the high-risk group and others into the low-risk group, 64 high risk patients (42.4%) had a multivariable-adjusted 5.15-fold (p < 0.001) increased risk of having higher Duke Index, compared with 87 patients (57.6%) at the low-risk group.

3.4. Internal and external validation of the GES12 for predicting coronary stenosis severity in patients with suspected CAD

For the internal validation cohort (n = 209), when the severity of coronary atherosclerosis was dichotomized based on the median values of modified Gensini scores and Sullivan Extent scores, the AUC for the GES12 was 0.798 (95%CI: 0.738–0.857) for detecting higher modified Gensini scores (> 31 *vs.* ≤ 31), and 0.776 (95%CI: 0.714–0.839) for predicting higher Sullivan Extent scores (> 30.5 *vs.* ≤ 30.5, Fig. 2D). Similar analyses in the external validation cohort (n = 206) yielded an AUC of 0.812 (95%CI: 0.754–0.870) for prediction of higher modified Gensini scores (> 27 *vs.* ≤ 27) and 0.778 (95%CI: 0.715–0.840) for prediction of higher Sullivan Extent scores (> 29 *vs.* ≤ 29, Fig. 2G). For both cohorts, calibration curves consistently showed a good agreement between prediction by the GES12 and the actual observation, suggesting that the GES12 model fitted well (p_{fit} > 0.05, Fig. 2E and H). When patients of two cohorts were categorized into GES12-defined high-risk and low-risk subgroups based on the cut point (i.e. 4.983) derived from the training set, patients in the high-risk group had significantly higher odds of having higher modified Gensini scores and higher Sullivan Extent scores compared with the low-risk group (Fig. 2F and I).

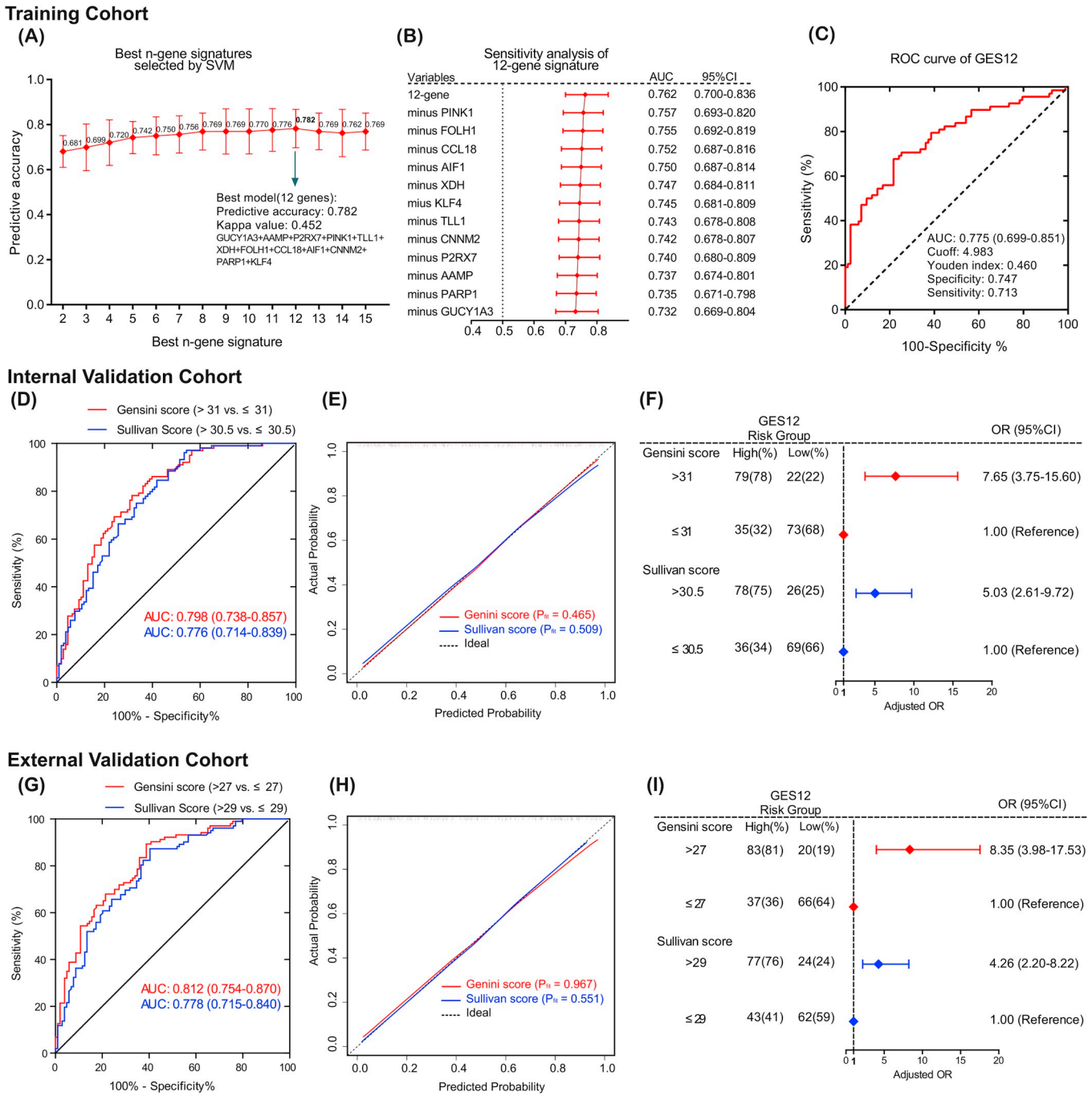

Fig. 2. Development and validation of GES12 for predicting the severity of coronary artery stenosis in training, internal validation, and external validation sets. (A) The predicted accuracy of the best n-gene signature constructed by 15 bioinformatically identified hub genes. (B) AUCs for the 12-gene signature and other twelve-minus-one gene signatures. (C) ROC curves of the GES12 for the prediction of higher Duke Index. (D and G) ROC curves of the GES12 for the prediction of higher modified Gensini scores and Sullivan Extent scores. (E and H) Calibration curves of the GES12. (F and I) Effect sizes for the associations of the GES12 with modified Gensini scores and Sullivan Extent scores after adjusting for age, sex, smoking status, alcohol drinking status, BMI, TG, TC, LDL-c, HDL-c, hypertension status, and diabetic status.

3.5. Diagnostic performance of the GES12 for obstructive CAD in validation cohorts

As presented in Fig. 3A and B, the GES12 showed good discrimination of obstructive CAD (defined as > 50% stenosis in ≥1 coronary vessel) from non-CAD subjects (internal validation: AUC = 0.834, 95%CI: 0.777–0.892; external validation: AUC = 0.792, 95%CI: 0.724–0.860). For both cohorts, no effect modification by cardiovascular risk factors was observed (Supplementary Table 7). When the reference standard for defining obstructive CAD was set at 55%–75% stenosis on coronary angiography, the diagnostic performance of the GES12, as determined by the AUC, was above 0.75; it slightly reduced to 0.71–0.72 only at an extreme standard of 80%–90% stenosis (Fig. 3C and D). In a vessel-based analysis that assessed the discriminative performance of the GES12 in each of three main coronary vessels, ROC curves showed similar AUCs for detecting stenosis

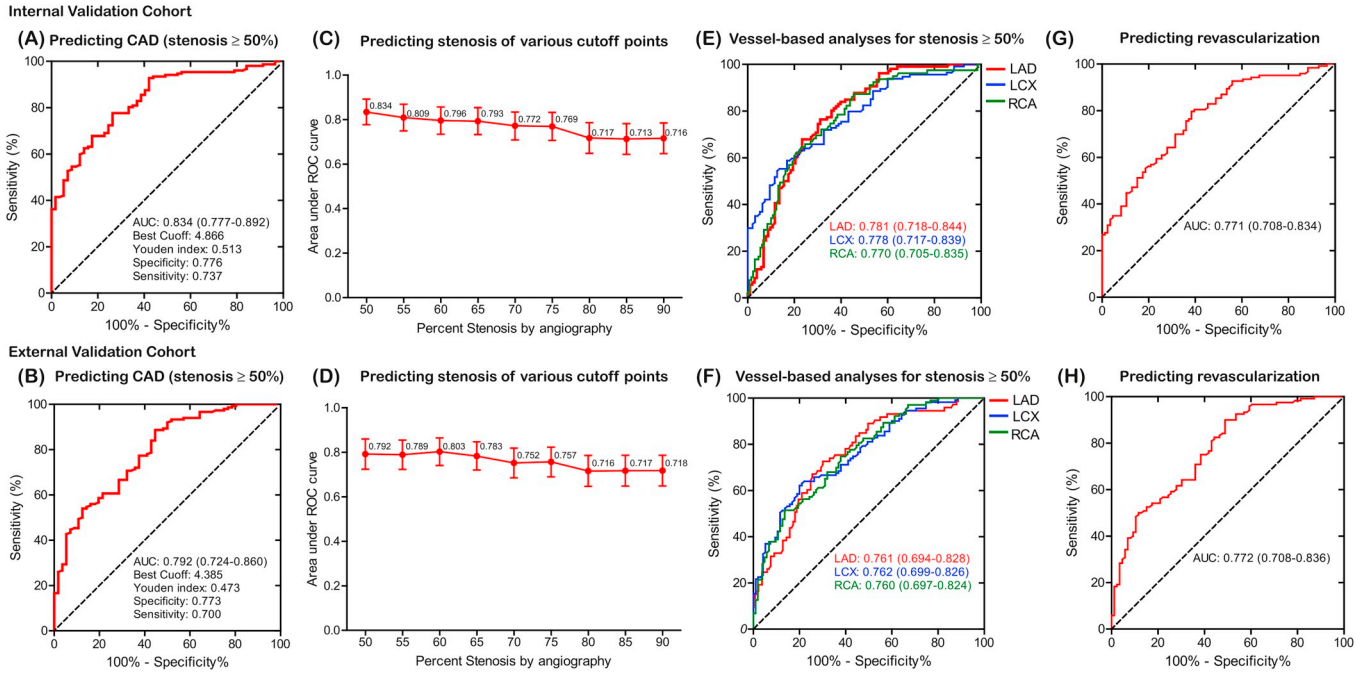

Fig. 3. Performance of GES12 in predicting obstructive CAD in internal and external validation sets.
(A and B) ROC curves assessing the predictive performance of GES12 to identify coronary stenosis in at least one vessel. (C and D) AUCs for predicting coronary stenosis at various cutoff points (50–90%) as measured by angiography. (E and F) ROC curves for GES12 to identify coronary stenosis of ≥50% in each of three coronary vessels. (G and H) ROC curves for GES12 to predict revascularization event within 30 days after diagnostic angiography.

of > 50% occurring at left anterior descending (including the left main coronary artery), left circumflex, and right coronary artery (Fig. 3E and F).

3.6. Predictive performance of the GES12 for revascularization events in two validation cohorts

During 30-day follow-up after angiography, ~58% of patients in two validation cohorts (internal: 58.9%; external: 58.3%, Table 1) underwent surgical or percutaneous revascularization. The AUCs of the GES12 for prediction of revascularization events were around 0.77 in two validation cohorts (internal: 0.771, 95%CI: 0.708–0.834; external: 0.772, 95%CI: 0.708–0.836, Fig. 3G and H).

3.7. Development of a nomogram including GES12 and traditional cardiovascular risk factors in the training set

To further increase the predictive performance of the GES12, we developed a clinically applicable nomogram (Fig. 4A), which integrated independent risk factors for higher Duke Index (Supplementary Table 8), including the GES12, smoking status, hypertension status, low density lipoprotein cholesterol (LDL-c) level, and body mass index (BMI). In the training set, the nomogram showed superiority in discrimination of higher Duke Index compared with the GES12 alone (0.839 *vs.* 0.775, $p_{\text{differnece}} = 0.022$, Fig. 4B). In decision curve analyses, the nomogram had larger net benefits in comparison with the GES12 across almost all the range of risk thresholds (Fig. 4C). Especially at a risk threshold range of 5–20%, the nomogram would identify 8-15 additional true events with higher Duke Index per 100 cases compared with the GES12. Classification performance of the nomogram and the GES12 was also compared. Overall, 57 (38%) of 151 patients were reclassified by the nomogram, with 41 (72%) reclassified correctly and only 16 (28%) reclassified incorrectly (categorical NRI = 0.317, IDI = 0.107, $p < 0.001$, Supplementary Table 9). In two validation cohorts, the nomogram also showed good performance in discrimination, clinical utility, and risk reclassification for predicting coronary artery stenosis (Supplementary Figs. 6 and 7 and Supplementary Table 10).

3.8. Associations between GES12 and plaque phenotypes in an IVUS substudy from the internal validation cohort

Considering that the functional significance of atherosclerotic lesions was not only governed by the degree of luminal narrowing, but also by plaque phenotypes [38], we further assessed the association of the GES12 with IVUS-defined plaque characteristics. A total of 202 lesions were identified by single-scale IVUS in 152 patients. On lesion-based analysis, the GES-defined high-risk group (GES12 > 4.983) had higher plaque burden ($p = 0.002$) and plaque volume ($p = 0.020$, Supplementary Table 11) than the low-risk group (≤4.983). On patient-based analysis, either the GES12 or the GES12-derived nomogram showed good discrimination of patients with VH-IVUS-defined thin-cap fibroatheromas (TCFA) lesions and lesions with a plaque burden of ≥70% (AUC = 0.72–0.75, Supplementary Fig. 8), suggesting a possible link between GES12 and markers of high-risk plaques [39].

4. Discussion

Extending previous studies that used WGCNA to identify specific modules or molecular subgroups of CAD [22,40,41], we aimed to screen hub genes of DM into a gene signature for prediction of coronary stenosis severity. We first identified one gene module significantly associated with the Duke Index in a whole blood gene expression profile extracted from the GEO database. Further enrichment analyses revealed significant over-representation of CAD-related categories in the DM. Then, by combining the results of WGCNA and PPI networks, we reduced the dimensionality of the module to a much smaller set of 15 hub genes that individually correlated with the Duke Index.

The major challenges for constructing a reproducible GES from a hybridization-based gene expression profile are lack of replication by a more stable technique and the risk of overfitting in predictive modeling of high-dimensional data [10]. To overcome these problems, we

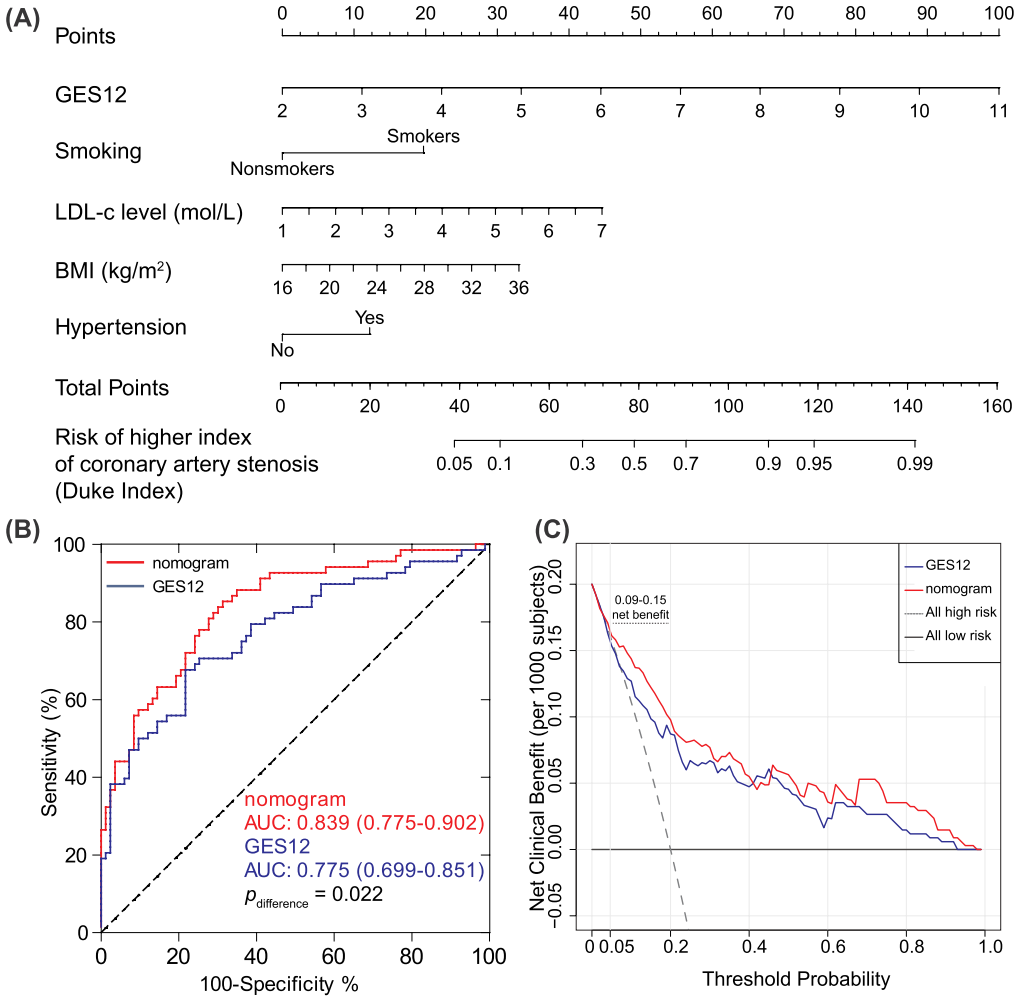

Fig. 4. Performance of the GES12-based nomogram in predicting higher Duke Index in the discovery phase.
(A) Developed GES12-based nomogram. (B) Comparative ROC analysis for comparing the discriminative ability between the nomogram and the GES12. (C) Decision curves analysis for comparing the clinical benefit between the nomogram and the GES12.

conducted a multi-step approach in a training set of 151 CAD patients, starting with RT-PCR replication of the 15 bioinformatically identified genes, followed by the use of a SVM algorithm for narrowing genes into a multi-gene panel. As a result, 12 of the 15 genes were selected to construct a 12-gene signature that provided the best predictive accuracy for higher Duke Index among all possible combinations of 15 genes. The overfitting issue was largely addressed by a sensitivity analysis (removing each gene at a time) showing that each of 12 genes was essential for the construction of a multi-gene signature. Finally, by weighting the expression levels of 12 genes with their effect sizes, we derived a GES12, which yielded good discrimination for higher Duke Index.

Considering that use of coronary angiographic scoring systems to estimate lesion severity is mainly challenged by the existence of heterogeneity among different scoring systems [19], we tried to validate the ability of the GES12 to reflect the degree of luminal narrowing, as determined by another two scoring systems, i.e. modified Gensini scores and Sullivan Extent scores. In patients with suspected CAD from two validation cohorts, the GES12 consistently showed a reliable accuracy (AUC: 76–84%) for predicting a broad spectrum of coronary artery stenosis, spanning from higher modified Gensini scores and Sullivan Extent scores to the presence of obstructive CAD to consequent revascularization events. Of note, patients in the GES12-defined high-risk group tended to have higher degree of luminal narrowing, and the low-risk group corresponded to the decreased lesion severity. Furthermore, by integrating the GES12 with four easily available cardiovascular risk factors (smoking status, hypertension status, LDL-c level, and BMI), we built a clinically practicable nomogram, which showed better performance, with respect to discrimination, risk classification, and clinical utility, for prediction of coronary artery stenosis than the GES12 alone. Taken together, our data suggest that the GES12, especially the GES12-based nomogram, may predict the severity of coronary artery stenosis in patients with known or suspect CAD.

Recently, Corus CAD, as a commercially available array incorporating age, sex, and 23 gene transcripts, has been introduced as an extremely promising predictor of obstructive CAD for non-diabetic patients [10]. In two multi-center studies, Corus CAD has yielded a large AUC (0.70–0.79), a high sensitivity (85–89%), but a relatively low specificity (43–52%) to discriminate ≥ 50% stenosis [13,42], suggesting that this array is more useful for CAD screening than for CAD diagnosis. In contrast, our newly established GES12, which showed a comparative AUC (0.792, 0.834) with a higher specificity (~77%) and an acceptable sensitivity (70–74%, Fig. 3) for the prediction of CAD, may be an additional choice when clinicians need to confidently treat suspected patients based on a positive test result. Also of note is that we did not observe effect modification by diabetic status on the predictive performance of the GES12 (Supplementary Table 7), suggesting that the GES12 may be also applicable for subjects with diabetes.

The advantage of a network approach like WGCNA is that it yields a contextual framework to explain how physically interactive genes

correlate with causal disease mechanisms [22]. For instance, *KLF4* and *PARP1*, as two gene transcripts of the GES12, have been experimentally identified to interact with each other to maintain telomere length [43], a determinant of atherosclerosis [44]. Another two contents of the GES12 (*GUCY1A3* and *CNNM2*) were also characterized by a gene-gene interaction analysis to constitute an interaction term associated with myocardial infarction and lipid processing [16]. From a biological perspective, inclusion of genes that may co-regulate in CAD-related pathways is a reasonable strategy to ensure the stable predictive performance of the GES12.

Limitations of our study included its retrospective design, although we enrolled one training set and two validation sets for rigorous construction of our signature. Next, our IVUS data suggested the GES12 as a potential predictor for high-risk plaques (i.e. TCFA lesions and plaque burden ≥70%). However, interpreting these results should be cautious, due to the relatively small sample size of the IVUS sub-study, potential sampling bias in single-vessel analyses, and lack of follow-up. Third, we consistently dichotomized the continuous severity scores at the median, in order to fully reflect the statistical distribution of severity scores in different studied cohorts. However, this also led to a problem that the cuts of severity scores were differing across cohorts, increasing the difficulty of interpreting the results. Fourth, the observation that no effect modification by cardiovascular risk factors on the performance of GES12 was possibly due to the decreased statistical power in subgroup analyses. Finally, it is of note that there were little overlapping genes between our GES12, Corus CAD, and a 9-gene signature identified by Muse et al. [12]. This phenomenon may be attributed to different strategies for gene signature selection, differences in disease phenotypes, or the complex nature of the pathogenesis of CAD leading to the coexistence of various genesets with certain predictive values [11].

In summary, the GES12, formulated by systematic use of gene network construction, SVM learning, and RT-PCR validation, is a promising predictor for the coronary stenosis severity in patients with known or suspected CAD. The nomogram, incorporating the GES12 and four cardiovascular risk factors, yields a 0-to-160 numeric scale allowing for personalized risk assessment of coronary artery stenosis. Prospective studies are needed to further validate the predictive performance of the GES12 and the GES12-based nomogram in multi-ethnic and diverse populations.

Financial support

This study was supported by grants from the National Basic Research Program of China (81800317) and the Union Program of the Key Scientific and Technological Project of Henan Province (2018020008).

Author contributions

XW and LM designed the experiments, wrote the paper, and draft the manuscript. NC and XL and performed the experiments. XW and XL interpreted the data.

Declaration of competing interest

The authors declared they do not have anything to disclose regarding conflict of interest with respect to this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atherosclerosis.2019.10.001.

References

[1] M. Naghavi, A.A. Abajobir, C. Abbafati, K.M. Abbas, F. Abd-Allah, S.F. Abera, et al., Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet 390 (10100) (2017) 1151–1210.
[2] J.M. Miller, C.E. Rochitte, M. Dewey, A. Arbab-Zadeh, H. Niinuma, I. Gottlieb, et al., Diagnostic performance of coronary angiography by 64-row CT, N. Engl. J. Med. 359 (22) (2008) 2324–2336.
[3] C. Ozcan, A. Deleskog, A.M. Schjerning Olsen, H. Nordahl Christensen, M. Lock Hansen, G. Hilmar Gislason, Coronary artery disease severity and long-term cardiovascular risk in patients with myocardial infarction: a Danish nationwide register-based cohort study, Eur. Heart J. Cardiovasc. Pharmacother. 4 (1) (2018) 25–35.
[4] M. Messerli, A.L. Panadero, A.A. Giannopoulos, M. Schwyzer, D.C. Benz, C. Grani, et al., Enhanced radiation exposure associated with anterior-posterior x-ray tube position in young women undergoing cardiac computed tomography, Am. Heart J. 215 (2019) 91–94.
[5] D. de Gonzalo-Calvo, D. Vilades, P. Martinez-Camblor, A. Vea, L. Nasarre, J. Sanchez Vega, et al., Circulating microRNAs in suspected stable coronary artery disease: a coronary computed tomography angiography study, J. Intern. Med. 286 (3) (2019) 341–355.
[6] J. Munkhaugen, J.E. Otterstad, T. Dammen, E. Gjertsen, T. Moum, E. Husebye, et al., The prevalence and predictors of elevated C-reactive protein after a coronary heart disease event, Eur. J. Prev. Cardiol. 25 (9) (2018) 923–931.
[7] M. Gaudino, F. Crea, Inflammation in coronary artery disease: which biomarker and which treatment? Eur. J. Prev. Cardiol. 26 (8) (2019) 869–871.
[8] J.S. Lin, C.V. Evans, E. Johnson, N. Redmond, E.L. Coppola, N. Smith, Nontraditional risk factors in cardiovascular disease risk assessment: updated evidence report and systematic review for the US preventive services task force, J. Am. Med. Assoc. 320 (3) (2018) 281–297.
[9] E. Ammirati, F. Fogacci, Clinical relevance of biomarkers for the identification of patients with carotid atherosclerotic plaque: potential role and limitations of cysteine protease legumain, Atherosclerosis 257 (2017) 248–249.
[10] M.A. Siemelink, T. Zeller, Biomarkers of coronary artery disease: the promise of the transcriptome, Curr. Cardiol. Rep. 16 (8) (2014) 513.
[11] B. Rhees, J.A. Wingrove, Developing peripheral blood gene expression-based diagnostic tests for coronary artery disease: a review, J. Cardiovasc. Transl. Res. 8 (6) (2015) 372–380.
[12] E.D. Muse, E.R. Kramer, H. Wang, P. Barrett, F. Parviz, M.A. Novotny, et al., A whole blood molecular signature for acute myocardial infarction, Sci. Rep. 7 (1) (2017) 12268.
[13] S. Rosenberg, M.R. Elashoff, P. Beineke, S.E. Daniels, J.A. Wingrove, W.G. Tingley, et al., Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients, Ann. Intern. Med. 153 (7) (2010) 425–434.
[14] J. Kim, N. Ghasemzadeh, D.J. Eapen, N.C. Chung, J.D. Storey, A.A. Quyyumi, et al., Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death, Genome Med. 6 (5) (2014) 40.
[15] J.A. Wingrove, S.E. Daniels, A.J. Sehnert, W. Tingley, M.R. Elashoff, S. Rosenberg, et al., Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis, Circ. Cardiovasc. Genet. 1 (1) (2008) 31–38.
[16] K. Hartmann, M. Seweryn, S.K. Handelman, G.A. Rempala, W. Sadee, Non-linear interactions between candidate genes of myocardial infarction revealed in mRNA expression profiles, BMC Genomics 17 (1) (2016) 738.
[17] P.R. Sinnaeve, M.P. Donahue, P. Grass, D. Seo, J. Vonderscher, S.D. Chibout, et al., Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease, PLoS One 4 (9) (2009) e7037.
[18] D.B. Mark, C.L. Nelson, R.M. Califf, F.E. Harrell Jr., K.L. Lee, R.H. Jones, et al., Continuing evolution of therapy for coronary artery disease. Initial results from the era of coronary angioplasty, Circulation 89 (5) (1994) 2015–2025.
[19] I.J. Neeland, R.S. Patel, P. Eshtehardi, S. Dhawan, M.C. McDaniel, S.T. Rab, et al., Coronary angiographic scoring systems: an evaluation of their equivalence and validity, Am. Heart J. 164 (4) (2012) 547–552 e541.
[20] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2) (2003) 249–264.
[21] J. Tang, D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong, et al., Prognostic genes of breast cancer identified by gene Co-expression network analysis, Front. Oncol. 8 (2018) 374.
[22] T. Huan, B. Zhang, Z. Wang, R. Joehanes, J. Zhu, A.D. Johnson, et al., A systems biology framework identifies molecular underpinnings of coronary heart disease, Arterioscler. Thromb. Vasc. Biol. 33 (6) (2013) 1427–1434.
[23] B. Vilne, J. Skogsberg, H. Foroughi Asl, H.A. Talukdar, T. Kessler, J.L.M. Bjorkegren, et al., Network analysis reveals a causal role of mitochondrial gene activity in atherosclerotic lesion formation, Atherosclerosis 267 (2017) 39–48.
[24] L. Yuan, G. Qian, L. Chen, C.L. Wu, H.C. Dan, Y. Xiao, et al., Co-expression network analysis of biomarkers for adrenocortical carcinoma, Front. Genet. 9 (2018) 328.
[25] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, et al., The comparative Toxicogenomics database: update 2019, Nucleic Acids Res. 47 (D1) (2019) D948–d954.
[26] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, et al., The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res. 45 (D1) (2017) D362–d368.
[27] G.Q. Zhu, Y. Yang, E.B. Chen, B. Wang, K. Xiao, S.M. Shi, et al., Development and validation of a new tumor-based gene signature predicting prognosis of HBV/HCV-included resected hepatocellular carcinoma patients, J. Transl. Med. 17 (1) (2019) 203.
[28] X.B. Wang, N.H. Cui, S. Zhang, S.R. Guo, Z.J. Liu, L. Ming, PARP-1 variant

Rs1136410 confers protection against coronary artery disease in a Chinese han population: a two-stage case-control study involving 5643 subjects, Front. Physiol. 8 (2017) 916.
[29] X.B. Wang, N.H. Cui, S. Zhang, Z.J. Liu, J.F. Ma, L. Ming, Leukocyte telomere length, mitochondrial DNA copy number, and coronary artery disease risk and severity: a two-stage case-control study of 3064 Chinese subjects, Atherosclerosis 284 (2019) 165–172.
[30] J. Hausleiter, T. Meyer, M. Hadamitzky, M. Zankl, P. Gerein, K. Dorrler, et al., Noninvasive coronary computed tomographic angiography for patients with suspected coronary artery disease: the Coronary Angiography by Computed Tomography with the Use of a Submillimeter resolution (CACTUS) trial, Eur. Heart J. 28 (24) (2007) 3034–3041.
[31] K. Chan, X. Pu, P. Sandesara, R.N. Poston, I.A. Simpson, A.A. Quyyumi, et al., Genetic variation at the ADAMTS7 locus is associated with reduced severity of coronary artery disease, J. Am. Heart Assoc. 6 (11) (2017).
[32] G. Niccoli, S. Giubilato, L. Di Vito, A. Leo, N. Cosentino, D. Pitocco, et al., Severity of coronary atherosclerosis in patients with a first acute coronary event: a diabetes paradox, Eur. Heart J. 34 (10) (2013) 729–741.
[33] S.P. Marso, A.D. Frutkin, S.K. Mehta, J.A. House, J.R. McCrary, V. Klauss, et al., Intravascular ultrasound measures of coronary atherosclerosis are associated with the Framingham risk score: an analysis from a global IVUS registry, EuroIntervention 5 (2) (2009) 212–218.
[34] L.C. Brumback, M.S. Pepe, T.A. Alonzo, Using the ROC curve for gauging treatment effect in clinical trials, Stat. Med. 25 (4) (2006) 575–590.
[35] K.M. Pencina, M.J. Pencina, R.B. D'Agostino Sr., What to expect from net reclassification improvement with three categories, Stat. Med. 33 (28) (2014) 4975–4987.
[36] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med. Decis. Mak. 26 (6) (2006) 565–574.
[37] M. Zhou, L. Hu, Z. Zhang, N. Wu, J. Sun, J. Su, Recurrence-associated long noncoding RNA signature for determining the risk of recurrence in patients with colon cancer, Mol. Ther. Nucleic Acids 12 (2018) 518–529.
[38] A.J. Brown, J.P. Giblett, M.R. Bennett, N.E.J. West, S.P. Hoole, Anatomical plaque and vessel characteristics are associated with hemodynamic indices including fractional flow reserve and coronary flow reserve: a prospective exploratory intravascular ultrasound analysis, Int. J. Cardiol. 248 (2017) 92–96.
[39] J.M. Cheng, H.M. Garcia-Garcia, S.P. de Boer, I. Kardys, J.H. Heo, K.M. Akkerhuis, et al., In vivo detection of high-risk coronary plaques by radiofrequency intravascular ultrasound and cardiovascular outcome: results of the ATHEROREMO-IVUS study, Eur. Heart J. 35 (10) (2014) 639–647.
[40] X. Zhang, R. Sun, L. Liu, Potentially critical roles of TNPO1, RAP1B, ZDHHC17, and PPM1B in the progression of coronary atherosclerosis through microarray data analysis, J. Cell. Biochem. 120 (3) (2019) 4301–4311.
[41] Peng, XY, Wang, Y and Hu, H. Identification of the molecular subgroups in coronary artery disease by gene expression profiles, J. Cell Physiol. (Published Online: February 25, 2019 (doi: 10.1002/jcp.28324)).
[42] G.S. Thomas, S. Voros, J.A. McPherson, A.J. Lansky, M.E. Winn, T.M. Bateman, et al., A blood-based gene expression test for obstructive coronary artery disease tested in symptomatic nondiabetic patients referred for myocardial perfusion imaging the COMPASS study, Circ. Cardiovasc. Genet. 6 (2) (2013) 154–162.
[43] M.H. Hsieh, Y.T. Chen, Y.T. Chen, Y.H. Lee, J. Lu, C.L. Chien, et al., PARP1 controls KLF4-mediated telomerase expression in stem cells and cancer cells, Nucleic Acids Res. 45 (18) (2017) 10492–10503.
[44] R.C. Stone, K. Horvath, J.D. Kark, E. Susser, S.A. Tishkoff, A. Aviv, Telomere length and the cancer-atherosclerosis trade-off, PLoS Genet. 12 (7) (2016) e1006144.