pISSN 1598-2998, eISSN 2005-9256

https://doi.org/10.4143/crt.2024.100

Original Article

A 10-Gene Signature to Predict the Prognosis of Early-Stage Triple-Negative Breast Cancer

Chang Min Kim[1,2], **Kyong Hwa Park**[3], **Yun Suk Yu**[1], **Ju Won Kim**[3], **Jin Young Park**[1], **Kyunghee Park**[4], **Jong-Han Yu**[5], **Jeong Eon Lee**[5], **Sung Hoon Sim**[6], **Bo Kyoung Seo**[7], **Jin Kyeoung Kim**[2], **Eun Sook Lee**[6], **Yeon Hee Park**[8], **Sun-Young Kong**[9,10]

[1]CbsBioscience. Inc., Daejeon, [2]Department of Pharmacy, College of Pharmacy, CHA University, Seongnam, [3]Division of Medical Oncology/Hematology, Department of Internal Medicine, Korea University College of Medicine, Seoul, [4]Samsung Genome Institute, Samsung Medical Center, Seoul, [5]Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, [6]Breast Cancer Center, National Cancer Center, Goyang, [7]Department of Radiology, Korea University Ansan Hospital, Korea University College of Medicine, Ansan, [8]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, [9]Targeted Therapy Branch, Research Institute, National Cancer Center, Goyang, [10]Department of Laboratory Medicine, Hospital, National Cancer Center, Goyang, Korea

Purpose Triple-negative breast cancer (TNBC) is a particularly challenging subtype of breast cancer, with a poorer prognosis compared to other subtypes. Unfortunately, unlike luminal-type cancers, there is no validated biomarker to predict the prognosis of patients with early-stage TNBC. Accurate biomarkers are needed to establish effective therapeutic strategies.

Materials and Methods In this study, we analyzed gene expression profiles of tumor samples from 184 TNBC patients (training cohort, n=76; validation cohort, n=108) using RNA sequencing.

Results By combining weighted gene expression, we identified a 10-gene signature (*DGKH*, *GADD45B*, *KLF7*, *LYST*, *NR6A1*, *PYCARD*, *ROBO1*, *SLC22A20P*, *SLC24A3*, and *SLC45A4*) that stratified patients by risk score with high sensitivity (92.31%), specificity (92.06%), and accuracy (92.11%) for invasive disease-free survival. The 10-gene signature was validated in a separate institution cohort and supported by meta-analysis for biological relevance to well-known driving pathways in TNBC. Furthermore, the 10-gene signature was the only independent factor for invasive disease-free survival in multivariate analysis when compared to other potential biomarkers of TNBC molecular subtypes and T-cell receptor β diversity. 10-gene signature also further categorized patients classified as molecular subtypes according to risk scores.

Conclusion Our novel findings may help address the prognostic challenges in TNBC and the 10-gene signature could serve as a novel biomarker for risk-based patient care.

Key words Triple-negative breast neoplasms, Biomarkers, Prognostic diagnosis, Gene signature

Introduction

Breast cancer is the most common cancer among women and is increasing globally, including in Korea [1-3]. Recent research has led to significant advances in new therapeutics and precision medicine, and active clinical applications of the novel approaches resulted in improved prognosis of patients with breast cancer [4-6].

Despite these advances, triple-negative breast cancer (TNBC) remains the most challenging subtype due to its aggressive and high metastatic potential even after a good response to standard systemic chemotherapy [7]. Since the first classification of TNBCs to six molecular subtypes based on mRNA expression, significant efforts have been made to elucidate subtypes of the highly heterogeneous cancer using new technologies [8-10]. Multi-omic analyses and new systematic approach were utilized for the better classification and more sophisticated therapeutic strategy in the recent studies [11-16]. Nonetheless, only a few biomarkers have been identified to accurately predict prognosis and guide treatment decisions [17].

In this study, we aimed to develop a clinically applicable prognostic gene signature for patients with TNBC using whole transcription sequencing to classify the highly heterogeneous cancer more accurately. We combined weighted genes related to prognosis and validated the results using

Correspondence: Sun-Young Kong
Targeted Therapy Branch, Research Institute, Department of Laboratory Medicine, Hospital, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Korea
Tel: 82-31-920-1735 Fax: 82-31-920-1379 E-mail: ksy@ncc.re.kr

Co-correspondence: Yeon Hee Park
Division of Hematology-Oncology, Department of Internal Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Korea
Tel: 82-2-3410-1780 Fax: 82-2-3410-1754 E-mail: yhparkhmo@skku.edu

Received January 29, 2024 **Accepted** May 6, 2024
Published Online May 10, 2024

*Chang Min Kim and Kyong Hwa Park contributed equally to this work.

statistical analysis and meta-analysis. We also validated the discovered gene signature in a cohort of patients with TNBC at another institution.

Materials and Methods

1. Patients and specimens

The study included 184 patients with early-stage TNBC; 76 patients in the training cohort were from the National Cancer Center Korea (NCC, Goyang, Korea), and 108 patients in the validation cohort were from Samsung Medical Center (SMC, Seoul, Korea). All patients were eligible if they were ≥ 18-years-old with early-stage TNBCs (stage I-III), for whom a histological biopsy could be safely obtained and standard systemic chemotherapy (anthracycline and taxane) and locoregional treatment including surgery and radiation were applied. Tumor samples were identified as TNBCs according to the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) guidelines [18,19]. The training cohort consisted of 15 patients who received neoadjuvant chemotherapy and 61 patients who underwent primary surgery followed by adjuvant chemotherapy for early-stage TNBC between March 2002 and August 2018. The validation cohort included 73 patients who had received neoadjuvant chemotherapy and 35 patients who received adjuvant chemotherapy after surgery between July 2011 and November 2017. In the validation cohort, 42 specimens of the neoadjuvant chemotherapy group were biopsy tissue before neoadjuvant chemotherapy, while the other specimens were surgical tissues. Tissues of training cohort were provided by NCC Bio Bank of National Cancer Center, Korea. All specimens were fresh-frozen.

2. Complete clinical information and outcome

Clinical data, including the date of diagnosis, clinical and surgical stages, response to neoadjuvant chemotherapy, recurrence, and survival, were collected from medical records. Invasive disease-free survival (iDFS) was defined as the time from diagnosis of primary breast cancer to invasive breast cancer recurrence or death from any cause and overall survival (OS) was defined as the time from diagnosis to death.

3. RNA sequencing and quality control

For the training cohort, sequencing libraries were prepared using fresh-frozen tissues with TruSeq Stranded mRNA LT Sample Prep Kit (Illumina Inc., San Diego, CA) following the manufacturer's protocols. paired-end sequencing was conducted using the prepared cDNA library for RNA sequencing using an Illumina HiSeq 4000 sequencer (Illumina). In RNA sequencing quality control, artifacts including adaptor sequences, contaminant DNA, and PCR duplicates were eliminated to reduce the bias of sequencing data. After quality control of sequencing data, aligned reads were generated by mapping sequencing data on the reference genome using the HISAT2 program (GitHub, http://daehwankimlab.github.io/hisat2/). With generated aligned reads, transcript assembly was conducted using StringTie (https://ccb.jhu.edu/software/stringtie/). Based on the transcript quantification of each sample, expression levels were normalized to transcript length and depth of coverage. Through normalization, expression profiles were extracted as fragments per kilobase of transcript per million mapped reads (FPKM).

For the validation cohort, sequencing libraries were prepared using fresh-frozen tissues with TruSeq RNA Sample Prep Kit v2 (Illumina Inc.) following the manufacturer's protocols. Sequencing of the RNA libraries was performed on a HiSeq 2500 sequencing platform (Illumina Inc.). After trimming poor-quality bases from the FASTQ files, we aligned the reads to the human reference genome (hg19) using STAR v.2.5 (GitHub, https://github.com/alexdobin/STAR) and estimated gene expression in terms of FPKM using RSEM v.1.3. The quality control of sequencing results was assessed using RNA-SeQC (v1.1.8).

For comparison between tumors and non-tumors, non-tumor data were collected from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). GSE58135 was selected as the non-tumor group in GEO, and it had non-tumor RNA sequencing data of 21 patients with TNBC [20]. In non-tumor data, data with a failed status or with the values of FPKM under 1.0×10^{-6} were excepted.

4. Differentially expressed gene and Cox regression analysis

By matching the genes of tumors and non-tumors, 10,856 genes were found in both groups. With 10,856 genes, the differentially expressed genes (DEGs) were analyzed. DEGs were screened to meet one of the following conditions: (1) statistically significant difference (Wilcoxon test, $|$fold change$|$ > 1.5, adjusted p-value < 0.05) comparing tumors with non-tumors and (2) statistically significant difference (Wilcoxon test, $|$fold change$|$ > 1.5, adjust p-value < 0.05) comparing the patients who exhibited recurrence/metastasis after surgical resection with the patients without recurrence/metastasis. The previously screened DEGs were further shortlisted, matching with the gene significant in Cox regression analysis for recurrence/metastasis. Before combining the DEGs, we identified the Cox regression coefficient of each gene and weighted gene expression with the corresponding coefficient value [21].

5. Combination gene analysis

The 59 genes selected in the DEG analysis were subjected to combination analysis ranging from 2 to 10 genes. The combinations were formed by multiplying the regression coefficients of each correlating gene with its expression level and then summing them. Continuous Cox regression analysis was then performed to obtain p-values (p-values saturated in combinations of 8 to 10 genes). Subsequently, receiver operating characteristic (ROC) statistical analysis was conducted on the combined gene sets to calculate the area under the curve (AUC).

6. Pre-validation of the candidate gene signatures for recurrence by cross validation of machine learning

Candidate gene signatures (achieving p-value < 0.05, AUC > 0.90, sensitivity > 80%, and specificity > 80%) were ranked by k-fold cross validation to identify the optimal gene combination. The patients were randomly separated by 2 folds (training set and test set) 300 times [21].

7. Signal transduction pathway analysis based on meta-analysis

Signal transduction analysis was performed using the CBS Probe PINGS (Reg. No. 2008-01-129-000568, CbsBioscience, Daejeon, Korea) [21]. For gene signature validation, signal transduction was analyzed for pathways related to each patient's DEGs compared recurrence/metastasis and non-recurrence/non-metastasis and for pathways related to gene signature. The genes were mapped to the signal transduction pathways obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg/). The top 10 signal transduction pathways were selected for each patient's DEGs and for gene signature according to the weight of the number of interactions and interacting genes. Ten pathways related to each patient's DEGs compared with GEO's non-tumor and gene signature-related pathways were compared. For each signal transduction pathway selected in signal transduction pathway analysis, the gene interaction frequency ratio was computed, which is a score of interacting genes with signature genes in gene signature validation. By applying 100% gene interaction frequency to the highest probability of gene interaction within each signal transduction pathway, the top 10 high interaction frequency genes were selected. In addition, 10 high interaction frequency genes related to each patient's DEGs and gene signatures related to high interaction frequency genes were compared.

8. Molecular subtype classification of TNBC

Molecular subtypes for both training and validation sets of data were primarily classified to six centroids using TNBCtype (http://cbc.mc.vanderbilt.edu/tnbc/) [22]. Using the results obtained from analysis, patients with TNBC were refined to four molecular subtypes such as basal-like 1 (BL1), basal-like 2 (BL2), mesenchymal (M), and luminal androgen receptor (LAR). Using Kaplan-Meier analysis, the prognostic power of TNBCtype-4 and the relationship between TNBCtype-4 and 10-gene signature were analyzed.

9. T-cell receptor diversity analysis

T-cell receptor (TCR) profiles were obtained using MiXCR 2.1.3 (GitHub, https://github.com/milaboratory/mixcr) using RNA sequencing data [23,24]. RNA sequencing data were aligned to all the IG/TCR loci. After two rounds of contig assembly, the V/J junctions of TCRs were extended. The assembled clonotypes were exported. TCR diversity was analyzed with the T cell receptor beta locus (TCRβ) using the Shannon index [25]. Using Kaplan-Meier survival (KM) analysis, the prognostic power of TCRβ diversity was analyzed.

10. Statistical analysis

Clinicopathological variables between the training and validation cohorts were evaluated using chi-square tests. Gene expression data were tested for normality using the Shapiro-Wilk test. As the data did not meet normality assumptions, significant differences between the responders and non-responders were evaluated using the Wilcoxon test. ROC curve analysis was used to determine the accuracy of threshold values for classifying recurrence/metastasis and non-recurrence/non-metastasis using gene signatures. KM curves were calculated using death and invasive disease as endpoints in iDFS and death in OS. The difference in KM curves was examined using the log-rank test, and the difference in hazard ratio was examined using Cox regression analysis. Candidate gene signatures were analyzed using Cox regression to understand the relationships between the recurrence/metastasis, classification, and clinicopathological variables. Significance was set at p < 0.05. All statistical analyses were performed using R v.3.4.3 software (R Development Core Team, https://www.r-project.org/).

Results

1. Clinical characteristics of patients

Of the 184 patients, 76 were in the training cohort and 108 in the validation cohort. The clinical characteristics of the patients in the training and validation cohorts are summarized in Table 1. Overall, the patients in the very young age group were not significantly different between the cohorts. Patients in the training cohort were more likely to have earlier-stage diseases than the validation cohort; however,

Table 1. Pathological baselines of the training and validation cohorts

Pathologic parameter	Training cohort			Validation cohort			Training vs. validation p-value[a]
	Total (n=76)	Primary tumors (n=61)	Residual tumors (n=15)	Total (n=108)	Primary tumors (n=77)	Residual tumors (n=31)	
Age (yr)							
< 35	17 (22.4)	10 (16.4)	7 (46.7)	36 (33.3)	24 (31.2)	12 (38.7)	0.1465
≥ 35	59 (77.6)	51 (83.6)	8 (53.3)	72 (66.7)	53 (68.8)	19 (61.3)	
TNM stage (pathologic)							
I	29 (38.2)	29 (47.5)	-	18 (16.7)	18 (23.4)	-	3.48×10^{-5}
II	42 (55.3)	32 (52.5)	10 (66.7)	57 (52.8)	44 (57.1)	13 (41.9)	
III	5 (6.6)	-	5 (33.3)	33 (30.6)	15 (19.5)	18 (58.1)	
Systemic chemotherapy							
Neo-adjuvant	15 (19.7)	-	15 (100)	73 (67.6)	42 (54.5)	31 (100)	-
AC	2 (13.3)	-	2 (13.3)	2 (2.7)	-	2 (6.5)	
AC-D	6 (40.0)	-	6 (40.0)	64 (87.7)	35 (83.3)	29 (93.5)	
AC-wP	-	-	-	7 (9.6)	7 (16.7)	-	
AC-PC	4 (26.7)	-	4 (26.7)	-	-	-	
PCarbo	2 (13.3)	-	2 (13.3)	-	-	-	
DA	1 (6.7)	-	1 (6.7)	-	-	-	
Adjuvant	61 (80.3)	61 (100)	-	35 (32.4)	35 (45.5)	-	
AC	6 (9.8)	6 (9.8)	-	2 (5.7)	2 (5.7)	-	
TC	14 (23.0)	14 (23.0)	-	1 (2.9)	1 (2.9)	-	
FAC	14 (23.0)	14 (23.0)	-	3 (8.6)	3 (8.6)	-	
AC-D	7 (11.5)	7 (11.5)	-	5 (14.3)	5 (14.3)	-	
AC-wP	6 (9.8)	6 (9.8)	-	4 (11.4)	4 (11.4)	-	
AC-PC	3 (4.9)	3 (4.9)	-	-	-	-	
TAC	11 (18.0)	11 (18.0)	-	20 (57.1)	20 (57.1)	-	
Event							
Presence	13 (17.1)	8 (13.1)	5 (33.3)	23 (21.3)	9 (11.7)	14 (45.2)	0.4057
Absence	63 (82.9)	53 (86.9)	10 (66.7)	85 (78.7)	68 (88.3)	17 (54.8)	
Specimen							
Core biopsy	-	-	-	42 (38.9)	42 (54.5)	-	
Surgery	76 (100)	61 (100)	15 (100)	66 (61.1)	35 (45.5)	31 (100)	
Follow-up period (mo), median (range)	51.5 (4.6-230.8)	51.1 (11.5-230.8)	57.8 (4.6-154.8)	58.3 (6.6-99.8)	59.0 (12.9-92.7)	56.6 (6.6-99.8)	

Values are presented as number (%) unless otherwise indicated. AC, adriamycin, cyclophosphamide; AC-D, AC followed by docetaxel; AC-PC, AC followed by paclitaxel and carboplatin; AC-wP, AC followed by weekly paclitaxel; DA, docetaxel and adriamycin; Event, recurrence or metastasis; FAC, 5-FU, adriamycin, cyclophosphamide; PCarbo, paclitaxel and carboplatin; TAC, taxotere, and AC; TC, taxotere and cyclophosphamide; TNM, tumor-node-metastasis (American Joint Committee on Cancer stage). [a]p-values were calculated using the chi-squared test.

the distribution of clinical stage among the patients was not different between the two groups. For adjuvant chemotherapy, the TAC (taxotere, adriamycin, and cyclophosphamide) regimen was used more in the validation cohort, probably because there were more advanced-stage patients. A schematic representation of the patients and samples is shown in S1 Fig.

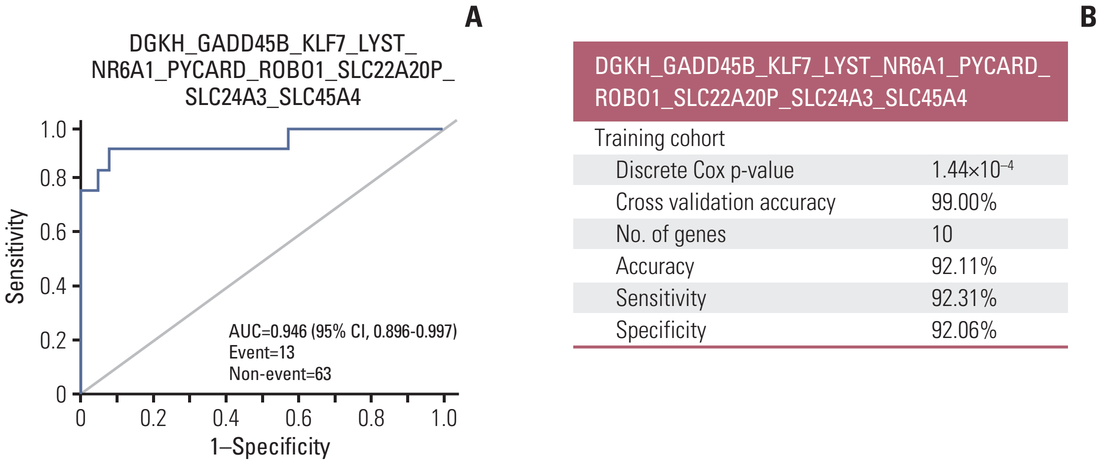
2. DEG analyses of tumor vs. non-tumor, and relapsed vs. non-relapsed

In the DEG analysis of tumor versus non-tumor, 9,741 genes were significantly differentially expressed by more than 1.5-fold changes in 10,856 genes. DEG analysis of primary tumors between relapsed and non-relapsed patients revealed 136 out of 10,856 genes showing significant differences in expression by 1.5-fold. Subsequently, out of 10,856 genes, 584 genes were statistically significant in the single

Table 2. Gene signature candidates as the prognostic biomarkers of triple-negative breast cancer

Rank	Gene signature	No. of combination genes	Continuous Cox p-value	AUC	Sensitivity	Specificity	Accuracy
1	DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ROBO1_SLC22A20P_SLC24A3_SLC45A4	10	1.24E-09	0.946	92.31	92.06	92.11
2	DGKH_DIP2B_EMP1_GADD45B_MT2A_NOTCH2_NR6A1_RORA_SLC22A20P_SLC24A3	10	1.35E-09	0.937	84.62	98.41	96.05
3	GADD45B_LYST_NOXA1_NR6A1_PYCARD_SLC22A20P_SLC24A3_NTAQ1	8	2.22E-09	0.941	84.62	96.83	94.74
4	DGKH_KLF7_LYST_NR6A1_ROBO1_SLC24A3_SLC6A20	7	2.32E-09	0.919	84.62	90.48	89.47
5	DGKH_EMP1_GADD45B_LYST_SLC22A20P_SLC24A3_SLC6A20	7	2.33E-09	0.950	92.31	88.89	89.47
6	CUEDC1_DGKH_EMP1_LYST_NOXA1_SLC22A20P_SLC6A20	7	2.76E-09	0.933	84.62	90.48	89.47
7	DGKH_KLF7_LYST_NR6A1_PRICKLE1_ROBO1_SLC24A3	7	2.97E-09	0.927	92.31	92.06	92.11
8	DCLK2_GADD45B_LYST_NR6A1_SLC22A20P_SLC24A3_NTAQ1	7	3.16E-09	0.954	84.62	98.41	96.05
9	GADD45B_KLF7_LYST_NR6A1_ROBO1_SLC22A20P_SLC6A20	7	3.33E-09	0.918	84.62	93.65	92.11
10	DCLK2_GADD45B_LYST_NR6A1_SLC22A20P_NTAQ1	6	5.09E-09	0.958	84.62	95.24	93.42

AUC, area under the curve; CUEDC1, CUE domain containing 1; DGKH, diacylglycerol kinase eta; DCLK2, doublecortin like kinase 2; DIP2B, disco interacting protein 2 homolog B; EMP1, epithelial membrane protein 1; GADD45B, growth arrest and DNA damage inducible beta; KLF7, Kruppel-like factor 7; LYST, lysosomal trafficking regulator; MT2A, metallothionein 2A; NOTCH2, notch receptor 2; NOXA1, NADPH oxidase activator 1; NR6A1, nuclear receptor subfamily 6 group A member 1; NTAQ1, N-terminal glutamine amidase 1; PRICKLE1, prickle planar cell polarity protein 1; PYCARD, PYD and CARD domain containing; ROBO1, roundabout guidance receptor 1; RORA, RAR related orphan receptor A; SLC22A20P, solute carrier family 22 member 20, pseudogene; SLC24A3, solute carrier family 24 member 3; SLC45A4, solute carrier family 45 member 4; SLC6A20, solute carrier family 6 member 20.



DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ROBO1_SLC22A20P_SLC24A3_SLC45A4	
Training cohort	
Discrete Cox p-value	1.44×10^{-4}
Cross validation accuracy	99.00%
No. of genes	10
Accuracy	92.11%
Sensitivity	92.31%
Specificity	92.06%

Fig. 1. Selected prognostic gene signature evaluation of clinical performance. The clinical performance of the prognostic gene signature was evaluated using receiver operating characteristic (ROC) analysis, cross validation, and Cox regression analysis. (A) ROC analysis of the prognostic gene signature to predict the recurrence of triple-negative breast cancer. (B) Clinical performance of the gene signature in Cox regression analysis, cross-validation, and ROC analysis. AUC, area under the curve; CI, confidence interval; DGKH, diacylglycerol kinase eta; GADD45B, growth arrest and DNA damage inducible beta; KLF7, Kruppel-like factor 7; LYST, lysosomal trafficking regulator; NR6A1, nuclear receptor subfamily 6 group A member 1; PYCARD, PYD and CARD domain containing; ROBO1, roundabout guidance receptor 1; SLC22A20P, solute carrier family 22 member 20, pseudogene; SLC24A3, solute carrier family 24 member 3; SLC45A4, solute carrier family 45 member 4.

Cox analysis. In the DEG analysis in three ways, 59 DEGs overlapped (S2 Table).

3. Candidate gene signatures by gene combination and selected gene signature by cross validation

The top 10 candidate gene signatures were ranked with the Continuous Cox p-value. Ten candidates showed values of 80 or higher in sensitivity, specificity, and accuracy. The prognostic gene signature was selected by meeting statistical criteria within subgroups of cohorts. Also the selected gene signature was *DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ROBO1_SLC22A20P_SLC24A3_SLC45A4*, showing 99.00% cross-validation accuracy, and it was statistically significant in the discrete Cox analysis. The risk score was calculated with a cut-off value of 5.959715 as follows: (0.818636×*DGKH*)+(0.018069×*GADD45B*)+(0.605352×*KLF7*) +(0.231666×*LYST*)+(1.305352×*NR6A1*)+(–0.052086×*PYCARD*)+(–0.196973×*ROBO1*)+(0.968759×*SLC22A20P*)+(0.098331×*SLC24A3)*+(0.311646×*SLC45A4*) (Table 2, Fig. 1).

4. Prognostic significance of gene signature in the training cohort

During the median follow-up of 51.5 months (range, 4.6 to 230.8 months), patients with tumors with high-risk gene signatures (n=17) showed significantly shorter iDFS (median, 58.5 months; 95% confidence interval [CI], 25.8 to not reached) than those with low-risk signatures (n=59, median not reached, p=1.32e-11) in the overall population (Fig. 2A). Further analysis in the separate group of patients who underwent primary surgery and in the patients with residual tumors after neoadjuvant chemotherapy showed similar results. Among patients who received primary surgery and patients with residual tumors after neoadjuvant chemotherapy, the median iDFS in the high-risk group was 68.9 months (95% CI, 58.5 to not reached; p=1.12e-05) and 25.8 months (95% CI, 10.6 to not reached; p=1.83e-05), respectively, and the median iDFS in the low-risk group did not reach the median (Fig. 2B and C).

The impact of gene signature on OS of training set represented significant difference in total patients (p=0.00019). However, subgroup analysis represented significant impact only in neoadjuvant treated patients subgroup (S3 Fig.).

5. Prognostic significance of gene signatures in the validation cohort

Median follow-up time for validation cohort was 58.3 months (range, 6.6 to 99.8 months). In the overall validation cohort, although the median iDFS of the patients with high-risk genetic signatures was not reached, there was a significantly higher risk for recurrence or metastasis than patients with low-risk gene signatures (median iDFS not reached and p=5.84e-06 in log-rank test). When the patients were sub-divided according to the treatment sequence, the prognostic significance of the gene signatures in the surgical tis-

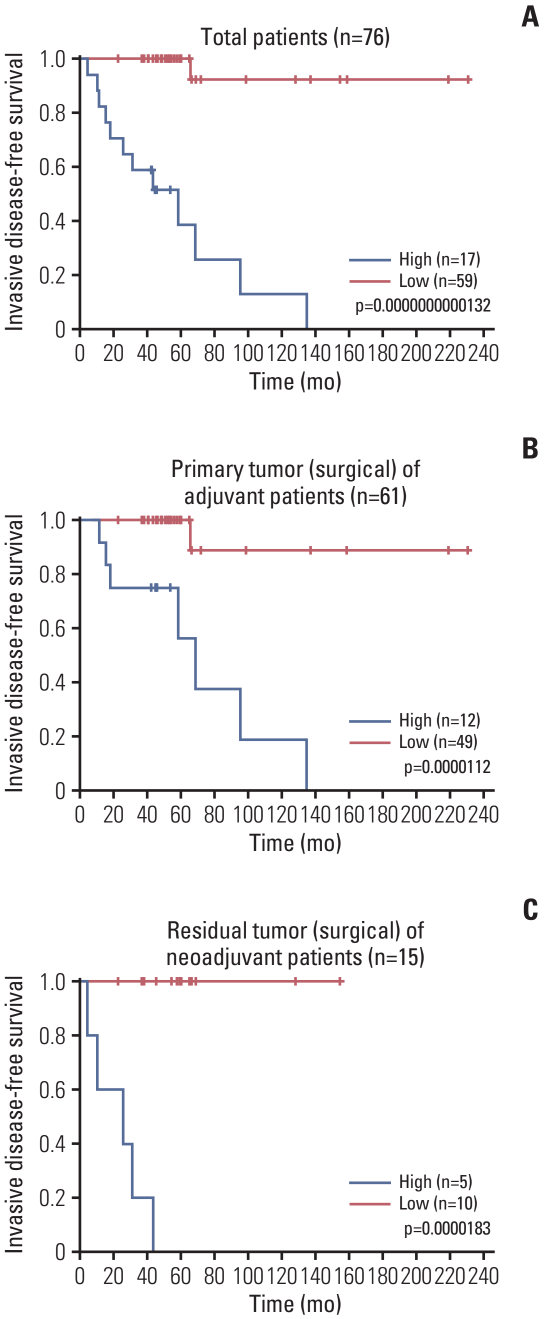


Fig. 2. Invasive disease-free survival in high risk and low-risk groups. The invasive disease-free survival was analyzed in different cases. (A) Kaplan-Meier curves for all patients. (B) Kaplan-Meier curves of patients treated with adjuvant chemotherapy. (C) Kaplan-Meier curves of patients treated with neoadjuvant chemotherapy.

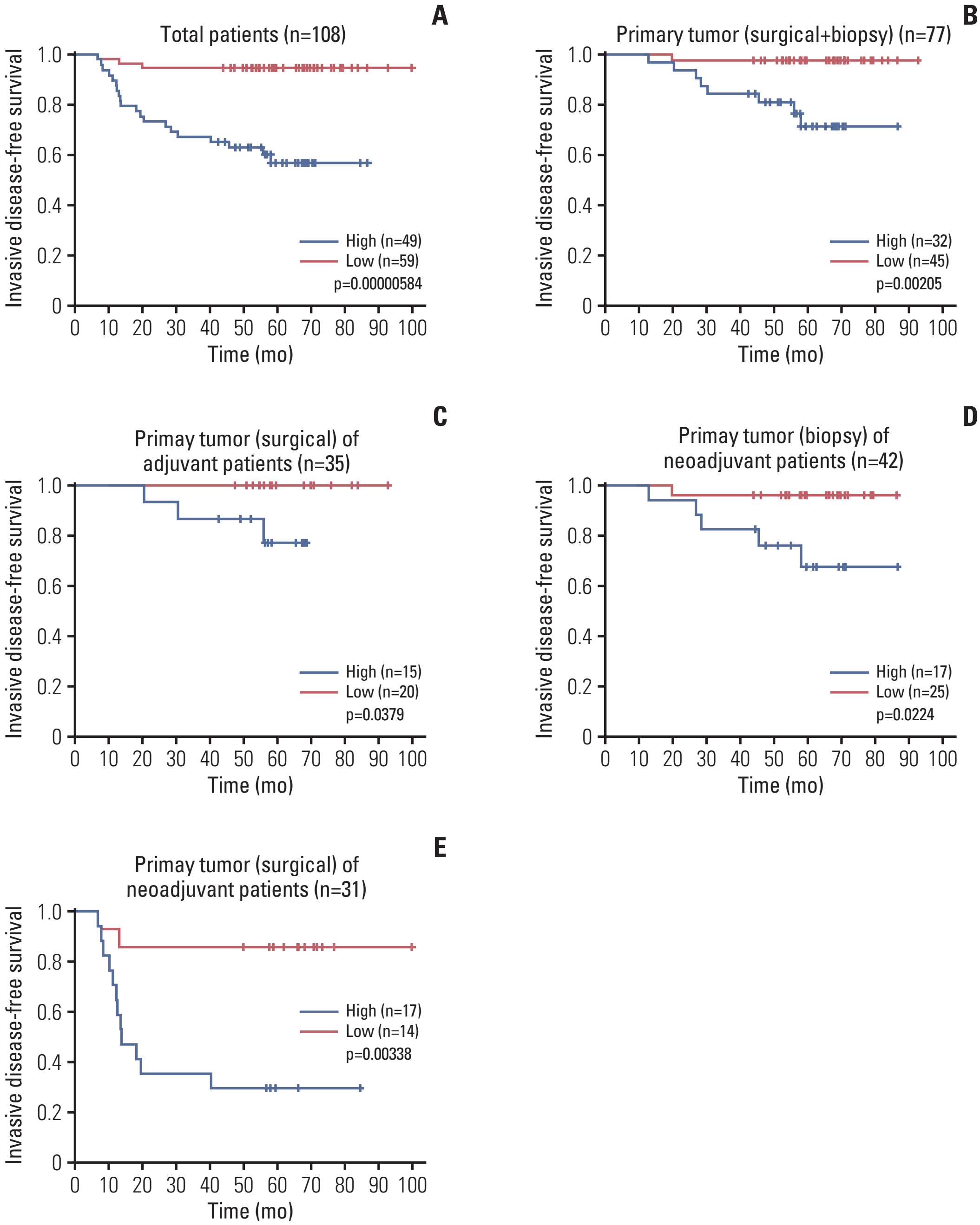
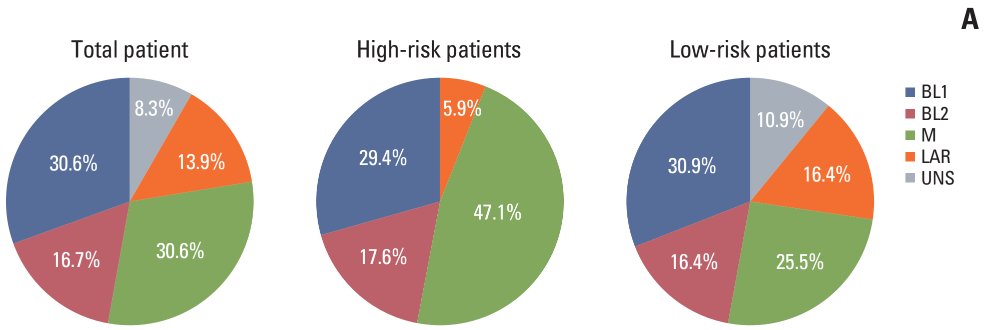


Fig. 3. Prognostic validation of the gene signature in the validation cohort. The prognostic gene signature was validated by invasive disease-free survival analysis in various cases. (A) Kaplan-Meier curves for all patients. (B) Kaplan-Meier curves in primary tumor specimens (surgical specimens of adjuvant patients and biopsy specimens of neoadjuvant patients). (C) Kaplan-Meier curves of patients treated with adjuvant chemotherapy. (D) Kaplan-Meier curves in biopsies of patients treated with neoadjuvant chemotherapy. (E) Kaplan-Meier curves for invasive disease-free survival in residual tumors of patients treated with neoadjuvant chemotherapy.

sues from the patients who underwent primary surgery was consistent with the training cohort (p=0.0379); however, the median iDFS in the high-risk and low-risk groups was not reached yet. High-risk gene signatures were still valid in predicting the prognosis of patients with residual tumors after neoadjuvant chemotherapy; the median iDFS in the high-risk group was 13.6 months (95% CI, 12.2 to not reached), but it was not reached in the low-risk group (p=3.38e-03).



TNBCtype	High-risk (%)	Low-risk (%)	Total (%)	P-value
BL1	5 (29.4)	17 (30.9)	22 (30.6)	0.3709
BL2	3 (17.3)	9 (16.4)	12 (16.7)	
M	8 (47.1)	14 (25.5)	22 (30.6)	
LAR	1 (5.9)	9 (16.4)	10 (13.9)	
UNS	0	6 (10.9)	6 (8.3)	
Total	17 (100)	55 (100)	72 (100)	

Fig. 4. TNBCtype-4 analysis with gene signature. TNBCtype-4 was analyzed by invasive disease-free survival analysis in various cases. (A) Pie charts of TNBCtype-4 for total patients, high-risk patients and low-risk patients. *(Continued to the next page)*

Also, when the gene signatures were examined in the tissues obtained by core biopsy in the neoadjuvant chemotherapy group, prognostic significance in iDFS was statistically significant (p=0.0224) (Fig. 3).

In validation set, the gene signature impact on OS represented significance in total patients (p=0.00183). However, subgroup analysis represented significant impact in only primary tumor groups (p=0.0190) and marginal significance in neoadjuvant-treated patients group (p=0.0510) (S4 Fig.).

6. Relationships between 10-gene signature and other potential prognostic factors

To investigate the interaction between the gene signature and other prognostic methods, specific prognostic values of the TNBCtype-4 were evaluated. In the TNBCtype-4 subgroup analysis, 76 patients with TNBC were classified into 22 patients of BL1 type (30.6%), 12 patients of BL2 type (16.7%), 22 patients of M type (30.6%), 10 patients of LAR type (13.9%) and six patients of unknown type (8.3%). There was no significant difference in terms of iDFS in the KM analysis by subtypes. However, patients could be further classified into high-risk and low-risk group by the gene signature in each subtype, respectively (Fig. 4). Among the TNBC subtypes, median iDFS of high-risk patients in 10-gene signature in BL1 (n=5, median iDFS was 15.5 months) and M (n=8, median iDFS was 58.5 months) subtypes were significantly inferior to those with low-risk patients (median not reached for both subtypes; p=3.41e-06 and p=4.29e-03 for BL1 and M, respectively). Additionally, we performed the TCRβ diversity analysis as a potential prognostic marker. With the cut-off that the highest point of the Youden index in ROC analysis (cut-off: 5.26), 35 patients belonged to the high diversity of TCRβ group and the rest of the patients had low diversity of TCRβ (n=41). However, the difference in iDFS according to the TCRβ diversity did not reach statistical significance (S5 Fig.).

7. Cox regression analysis of the selected gene signature

The prognostic impact of the selected gene signature (*DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ROBO1_SLC22A20P_SLC24A3_SLC45A4*) was investigated with Cox regression analysis. In the univariate Cox regression analysis, the gene signature was the most significant factor among the clinical factors and high score was strongly correlated with poor prognosis. TNM stage showed a marginal significance. In the multivariable Cox regression analysis with the gene signature, TNM stage, and TCRβ diversity, the gene signature only remained statistically significant (Table 3).

8. Signal transduction pathway analysis and high interaction frequency genes analysis for prognostic gene signature

Through the biological meta-analysis, we found gene signature and prognosis-related KEGG signal transduction pathways as well as high interaction frequency genes.

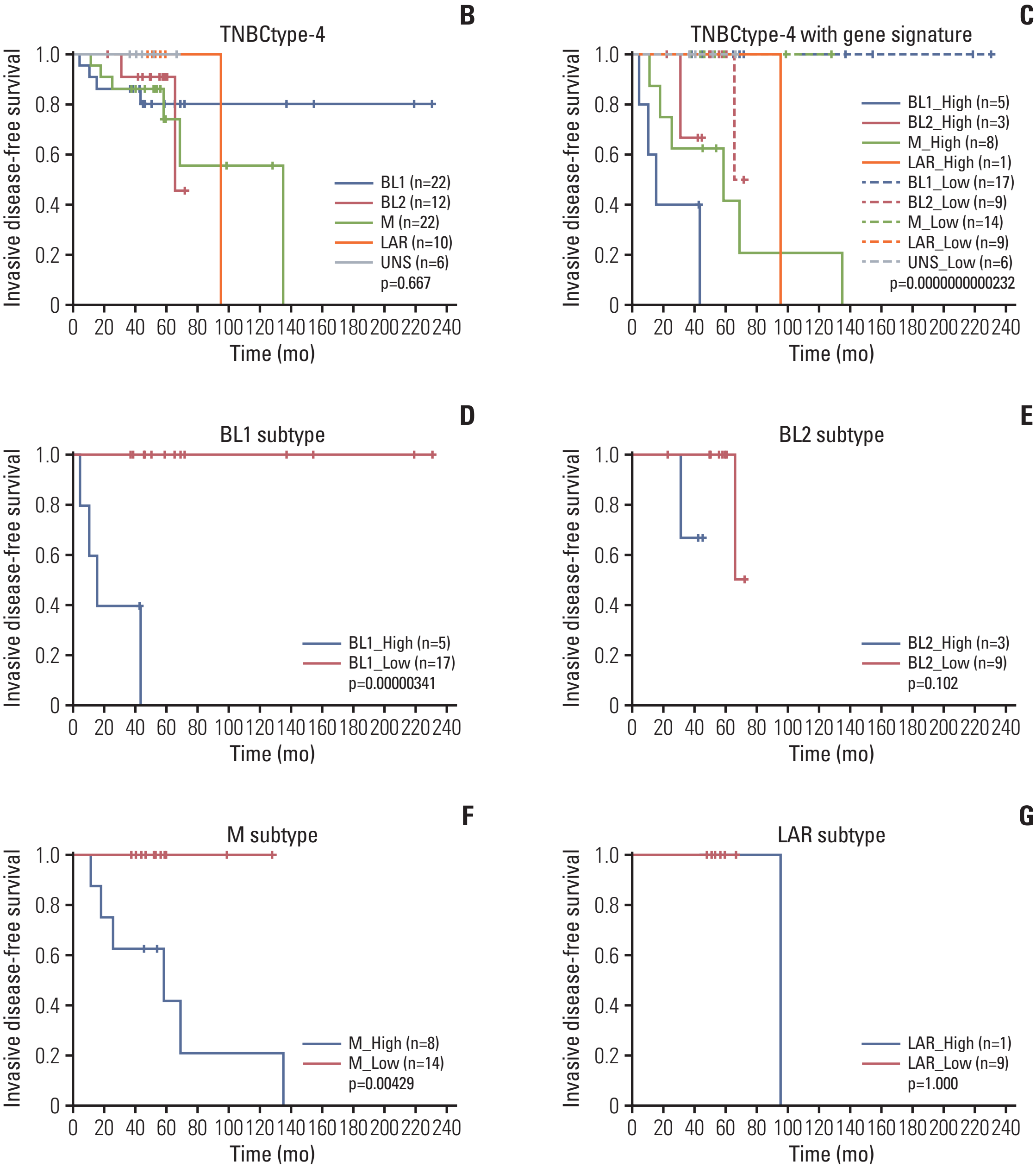


Fig. 4. *(Continued from the previous page)* (B) Kaplan-Meier curves of patients classified with TNBCtype-4. (C) Kaplan-Meier curves of patients classified with TNBCtype-4 and gene signature. (D) Kaplan-Meier curves of patients classified into BL1 subtype. (E) Kaplan-Meier curves of patients classified into BL2 subtype. (F) Kaplan-Meier curves of patients classified into M subtype. (G) Kaplan-Meier curves of patients classified into LAR subtype. BL1, basal-like 1; BL2, basal-like 2; LAR, luminal androgen receptor; M, mesenchymal; UNS, unspecified.

Transduction pathway analysis revealed that the pathways in cancer, phosphoinositide 3-kinase–Akt signaling pathway, Alzheimer disease pathway, Human cytomegalovirus infection pathway, hepatitis C pathway, breast cancer pathway, and mitogen-activated protein kinase signaling pathway were related to the prognostic gene signatures and prognosis. In these pathways, *KRAS*, *HRAS*, and *APP* were the high interaction frequency genes related to gene signatures and prognosis (S6 Table).

Table 3. Cox regression analysis of the prognostic gene signature and variables

Variable	No.	RC	HR (95% CI)	p-value
Univariable Cox regression analysis				
Candidate genes				
DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ ROBO1_SLC22A20P_SLC24A3_SLC45A4 (low vs. high)	76	3.9606	52.49 (6.81-404.53)	1.44×10^{-4}
Clinicopathological features				
Age (≤ 35 yr vs. > 35 yr)	76	–0.1605	0.85 (0.25-2.94)	0.799
TNM stage (1, 2 vs. 3)	76	0.9281	2.53 (0.55-11.74)	0.236
TNM stage (1 vs. 2, 3)	76	1.9985	7.38 (0.95-57.08)	0.056
PAM50 call ROR-S (subtype) (low, med, high)	76	–0.4875	0.61 (0.30-1.27)	0.186
TCRβ Diversity (low vs. high)	76	–0.8453	0.43 (0.13-1.41)	0.162
Multivariable Cox regression analysis				
DGKH_GADD45B_KLF7_LYST_NR6A1_PYCARD_ ROBO1_SLC22A20P_SLC24A3_SLC45A4 (low vs. high)	76	4.2879	72.81 (7.04-752.75)	3.21×10^{-4}
TNM stage (pathologic, 1 vs. 2, 3)	76	2.3537	10.52 (0.67-165.09)	0.094
TCRβ diversity (low vs. high)	76	0.3597	1.43 (0.37-5.56)	0.603

CI, confidence interval; DGKH, diacylglycerol kinase eta; GADD45B, growth arrest and DNA damage inducible beta; HR, hazard ratio; KLF7, Kruppel-like factor 7; LYST, lysosomal trafficking regulator; NR6A1, nuclear receptor subfamily 6 group A member 1; PYCARD, PYD and CARD domain containing; RC, regression coefficient; ROBO1, roundabout guidance receptor 1; ROR-S, risk of recurrence based on subtype; SLC22A20P, solute carrier family 22 member 20, pseudogene; SLC24A3, solute carrier family 24 member 3; SLC45A4, solute carrier family 45 member 4; TCRβ, T cell receptor beta locus; TNM, tumor-node-metastasis (American Joint Committee on Cancer stage).

Discussion

TNBCs are highly heterogeneous, making it challenging to predict prognosis and select appropriate treatments. Molecular categorization based on the comprehensive genetic signatures classified into distinct subtypes and therapeutic targets were suggested [14,26]. However, for the early-stage TNBCs, the introduction of immunotherapy was the only translation.

In this study, we aimed to identify and validate a 10-gene signature set based on the transcriptome of primary tumors for predicting the prognosis of patients with early-stage TNBC. Our analysis revealed that the gene signature set could better stratify patients with TNBC based on risk score than clinicopathologic parameters with an accuracy of 92.11% at a cut-off of 5.959715. We also validated the gene signature set was validated in patients with TNBC from another institution, demonstrating its objectivity.

Among the genes that make up the 10-gene set, some are related to immune responses (*LYST* and *PYCARD*), while others are related to the biology of cancer cells (*GADD45B*, *KLF7*, *ROBO1*, *SLC45A4*). The rest are novel genes with little evidence of a link to cancer (*DGKH*, *NR6A1*, *SLC22A20P*, *SLC24A3*). Detailed information on each gene that makes up the 10-gene set is presented in the S7 Table.

The objective of this study was to verify the 10-gene signatures in multivariate analysis by exploring additional biomarkers, in addition to traditional clinical variables. When our patient samples were analyzed according to the TNBC-type-4 subtyping [22], we observed a similar pattern of prognosis in our training cohort, based on subtypes (Fig. 4B). Further categorization of patients using the 10-gene signature revealed that patients with BL1 subtype and M subtype, but not BL2 and LAR subtypes, had a distinctly different prognosis (Fig. 4C-G). Using the gene score, we were able to differentiate patients with poor prognosis from those predicted to have a good prognosis for BL1 (5 out of 22, 22.7%). Among the total of 22 patients with the M subtype predicted to have poor risk, only eight (36.3%) were classified as high risk. However, the TCRβ diversity analysis showed that the mean level of TCRβ diversity was not significantly different between patients with relapsed and non-relapsed patients, and although patients with high TCRβ diversity (cut-off: 5.26) showed a trend for better iDFS, the difference was not statistically significant (S5 Fig.).

Interestingly, in the analysis using surgical specimens, the 10-gene signature discovered in this study significantly predicted the prognosis of patients with early-stage TNBC, regardless of whether the specimen was obtained from primary surgery or a residual tumor after neoadjuvant chemotherapy. In a recent analysis, 56% of cases showed a subtype change after neoadjuvant chemotherapy in patients with TNBC, and the most common change was from BL1 to M subtype [27]. Given that the BL1 and M subtypes account for the majority (55%-60%) of TNBC cases, the good performance of the 10-gene signature in these subtypes might have contributed to the consistent results in both pre-treatment and residual tumors.

Our approach to biomarker discovery differed from others in that we selected gene combinations that showed the best performance using combination analysis, which is a way to find optimal genes that can be used as biomarkers in large-scale analyses such as RNA sequencing. We then performed cross-validation as a pre-validation using a machine learning process. Through these procedures, we discovered a list of gene combinations with the least chance of failure in a separate validation cohort. Finally, a meta-analysis of the gene signature enabled us to demonstrate that the 10-gene signature has biological relevance in TNBC, not just as a list of genes with statistical power (S8 Fig.) [21].

We discovered a novel, 10-gene signature and validated it in a separate cohort of prospectively collected samples with regular follow-up data from a separate institution. There are several limitations in the application of the findings from this study. One of these limitations pertains to the heterogeneity of systemic chemotherapy regimens utilized, which were based on various clinical risk factors. It's important to note that the patients involved in this study were enrolled prior to the implementation of findings from the CREATE-X or Keynote-522 studies into clinical practice [28,29]. Consequently, further investigation is warranted to assess the predictive potential of the gene signature concerning adjuvant capecitabine and/or immunotherapy. Given that the current standard treatment for early-stage TNBC involves a combination of pembrolizumab and chemotherapy as neoadjuvant therapy, additional research is necessary to ascertain the predictive role of the 10-gene signature in achieving pathologic complete remission or guidance for escalated treatment in high-risk patients.

In conclusion, our study identified a 10-gene signature as a potential prognostic biomarker for patients with early-stage TNBC. To be used as a biomarker in a risk-based approach to clinical practice, this gene signature should be further validated in prospective clinical studies involving new treatments such as capecitabine and immune checkpoint inhibitors are applied. Nonetheless, we suggest that our findings can contribute to solving diagnostic challenges in TNBC and to a step close to precision medicine for qualified patient care.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (https://www.e-crt.org).

Ethical Statement

All patients provided written informed consent, and the study protocol was approved by the institutional review boards of the National Cancer Center Korea and Samsung Medical Center (NCC IRB # 2017-0127, 2019-0004, SMC IRB# 2014-11-015).

Author Contributions

Conceived and designed the analysis: Park KH, Park JY, Lee ES, Park YH, Kong SY.
Collected the data: Park KH, Park YH, Kong SY.
Contributed data or analysis tools: Kim CM, Park KH, Yu YS, Park YH, Kong SY.
Performed the analysis: Kim CM, Park KH, Yu YS, Kim JW, Park K, Yu JH, Lee JE, Park YH, Kong SY.
Wrote the paper: Kim CM, Park KH, Yu YS, Sim SH, Seo BK, Kim JK, Park YH, Kong SY.
Obtained funding: Park KH, Park JY, Park YH, Kong SY.
Administrative, technical, or material support: Park K, Park KH, Park YH, Lee ES, Kong SY, Park JY.

ORCID iDs

Chang Min Kim: https://orcid.org/0000-0002-5118-6031
Kyong Hwa Park: https://orcid.org/0000-0002-2464-7920
Yeon Hee Park: https://orcid.org/0000-0003-4156-9212
Sun-Young Kong: https://orcid.org/0000-0003-0620-4058

Conflicts of Interest

Chang Min Kim; Employee for CbsBioscience. Kyong Hwa Park; Consultancy/advisory role for: AstraZeneca, Pfizer, Eisai, Roche, Daiichi-Sankyo, Eli Lilly, and Novartis Pharmaceuticals; Honoraria from: AstraZeneca, Pfizer, Eisai, Roche, Daiichi-Sankyo, iMBDx, and Novartis; Grant/Research funding from AstraZeneca. Yun Suk Yu; Employee for CbsBioscience. Jin Young Park; Employer of CbsBioscience. Yeon Hee Park; Consultancy/advisory role for: AstraZeneca, Pfizer, Eisai, Roche, Daiichi-Sankyo, Eli Lilly, and Novartis Pharmaceuticals; Patents and royalties from: Hanmi; Honoraria from: AstraZeneca, Pfizer, Eisai, Roche, Daiichi-Sankyo, and Novartis; Grant/Research funding from AstraZeneca, Merck, Pfizer, Novartis, Alteogen, and Roche. Sun-Young Kong; Grant/Research funding from GSK and iMBDx. All remaining authors have declared no conflicts of interest.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (HI17C2206). This study was supported by the National R&D Program for Cancer Control, Ministry of Health and Welfare (HA15C0010) and National Cancer Center (#2210543), Republic of Korea.

References

1. Global Burden of Disease Cancer Collaboration; Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the Global Burden of Disease Study. JAMA Oncol. 2018;4:1553-68.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209-49.
3. Hong S, Won YJ, Park YR, Jung KW, Kong HJ, Lee ES, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2017. Cancer Res Treat. 2020;52:335-50.
4. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, et al. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet. 2018; 391:1023-75.
5. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N Engl J Med. 2018;379:111-21.
6. Henry NL, Somerfield MR, Abramson VG, Ismaila N, Allison KH, Anders CK, et al. Role of patient and disease factors in adjuvant systemic therapy decision making for early-stage, operable breast cancer: update of the ASCO endorsement of the cancer care Ontario guideline. J Clin Oncol. 2019;37:1965-77.
7. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res. 2007; 13:2329-34.
8. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res. 2015;21:1688-98.
9. Liu YR, Jiang YZ, Xu XE, Yu KD, Jin X, Hu X, et al. Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer. Breast Cancer Res. 2016;18:33.
10. Garrido-Castro AC, Lin NU, Polyak K. Insights into molecular classifications of triple-negative breast cancer: improving patient selection for treatment. Cancer Discov. 2019;9:176-98.
11. Tutt A, Tovey H, Cheang MC, Kernaghan S, Kilburn L, Gazinska P, et al. Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial. Nat Med. 2018;24:628-37.
12. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68.
13. Kim J, Yu D, Kwon Y, Lee KS, Sim SH, Kong SY, et al. Genomic characteristics of triple-negative breast cancer nominate molecular subtypes that predict chemotherapy response. Mol Cancer Res. 2020;18:253-63.
14. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121:2750-67.
15. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747-52.
16. Park J, Tacam MJ, Chauhan G, Cohen EN, Gagliardi M, Iles LR, et al. Nonphosphorylatable PEA15 mutant inhibits epithelial-mesenchymal transition in triple-negative breast cancer partly through the regulation of IL-8 expression. Breast Cancer Res Treat. 2021;189:333-45.
17. Sukumar J, Gast K, Quiroga D, Lustberg M, Williams N. Triple-negative breast cancer: promising prognostic biomarkers currently in development. Expert Rev Anticancer Ther. 2021; 21:135-48.
18. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol. 2013;31:3997-4013.
19. Allison KH, Hammond ME, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists guideline update. Arch Pathol Lab Med. 2020;144:545-63.
20. Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK, et al. Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res Treat. 2014;146:287-97.
21. Park IJ, Yu YS, Mustafa B, Park JY, Seo YB, Kim GD, et al. A nine-gene signature for predicting the response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer. Cancers (Basel). 2020;12:800.
22. Lehmann BD, Jovanovic B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. PLoS One. 2016;11:e0157368.
23. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017;35:908-11.
24. Han J, Duan J, Bai H, Wang Y, Wan R, Wang X, et al. TCR repertoire diversity of peripheral PD-1(+)CD8(+) T cells predicts clinical outcomes after immunotherapy in patients with non-small cell lung cancer. Cancer Immunol Res. 2020;8:146-54.
25. Cui JH, Lin KR, Yuan SH, Jin YB, Chen XP, Su XK, et al. TCR repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervical cancer. Front Immunol. 2018;9:2729.
26. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, et al. Comprehensive genomic analysis identifies

novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res. 2015;21:1688-98.

27. Masuda H, Harano K, Miura S, Wang Y, Hirota Y, Harada O, et al. Changes in triple-negative breast cancer molecular subtypes in patients without pathologic complete response after neoadjuvant systemic chemotherapy. JCO Precis Oncol. 2022; 6:e2000368.

28. Masuda N, Lee SJ, Ohtani S, Im YH, Lee ES, Yokota I, et al. Adjuvant capecitabine for breast cancer after preoperative chemotherapy. N Engl J Med. 2017;376:2147-59.

29. Schmid P, Cortes J, Pusztai L, McArthur H, Kummel S, Bergh J, et al. Pembrolizumab for early triple-negative breast cancer. N Engl J Med. 2020;382:810-21.