Carcinogenesis, 2020, Vol. 41, No. 7, 887–893

doi:10.1093/carcin/bgaa035
Advance Access Publication April 8, 2020
Original Article

ORIGINAL ARTICLE

Differences in gene-expression profiles in breast cancer between African and European-ancestry women

Jie Ping, Xingyi Guo, Fei Ye[1], Jirong Long, Loren Lipworth, Qiuyin Cai, William Blot, Xiao-Ou Shu and Wei Zheng

Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center and [1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

*To whom correspondence should be addressed. Tel: +615 936 0682; Fax: +615 936 8241; Email: wei.zheng@vanderbilt.edu

Abstract

African American (AA) women have an excess breast cancer mortality than European American (EA) women. To investigate the contribution of tumor biology to this survival health disparity, we compared gene expression profiles in breast tumors using RNA sequencing data derived from 260 AA and 155 EA women who were prospectively enrolled in the Southern Community Cohort Study (SCCS) and developed breast cancer during follow-up. We identified 59 genes (54 protein-coding genes and 5 long intergenic non-coding RNAs) that were expressed differently between EA and AA at a stringent false discovery rate (FDR) < 0.01. A gene signature was derived with these 59 genes and externally validated using the publicly available Cancer Genome Atlas (TCGA) data from180 AA and 838 EA breast cancer patients. Applying C-statistics, we found that this 59-gene signature has a high discriminative ability in distinguishing AA and EA breast cancer patients in the TCGA dataset (C-index = 0.81). These findings may provide new insight into tumor biological differences and the causes of the survival disparity between AA and EA breast cancer patients.

Introduction

Breast cancer is the most common cancer in women worldwide, with more than 2 million new cases diagnosed annually (1). Due to advances in early detection and improvements in treatment, breast cancer survival has increased substantially over the last three decades, particularly in developed countries where the 5-year survival rates for invasive breast cancer have reached 90% and higher (2). However, there is a significant racial disparity in breast cancer survival, with African American (AA) patients having a higher breast cancer mortality rate than European American (EA) patients. This disparity could be partially attributed to differences in breast cancer biology between AA and EA (3).

Several previous studies have profiled gene expression in breast cancer tissues to investigate possible distinct biological mechanisms and molecular pathways of breast carcinogenesis and prognosis between AA and EA women (4–8). A recent study analyzed racial differences in the expression levels of 200 genes, including 50 genes from the PAM50 panel, in breast cancer tissues collected from 495 AA and 478 EA patients included in the Carolina Breast Cancer Study (7). In another recent study, RNA sequencing (RNA-Seq) data generated from the breast cancer tissues of 154 AA and 774 EA breast cancer patients included in the Cancer Genome Atlas (TCGA) were analyzed (4). Both studies identified genes that were differentially expressed between AA and EA patients. However, none of these studies rigorously cross-validated their findings using an independent dataset, and thus some of the reported observations could be spurious due to type I errors.

In this study, we used RNA-Seq data generated from a large collection of breast tumors—from 260 AA and 155 EA patients from the Southern Community Cohort Study (SCCS)—to identify a gene signature that best differentiates between these two patient groups. We constructed a 59-genes signature using the SCCS data and externally validated this gene signature with

Received: February 18, 2020; **Revised:** March 25, 2020; **Accepted:** April 7, 2020

Abbreviations

AA	African American
EA	European American
ER	estrogen receptor
FDR	false discovery rate
FFPE	formalin-fixed paraffin-embedded
GO	Gene Ontology
HER2	human epidermal growth factor receptor 2
PR	progesterone receptor
RNA-Seq	RNA sequencing
SCCS	Southern Community Cohort Study
TCGA	the Cancer Genome Atlas
TNBC	triple negative breast cancer.

TCGA data by showing that this signature discriminates between tumors from AA and EA patients. Furthermore, we evaluated the association of the genes included in this signature with overall survival among AA or EA breast cancer patients. None of these genes, however, were significantly associated with survival after adjusting for multiple comparisons.

Methods

Study population

Breast cancer patients included in the current project are from the SCCS, a population-based, prospective cohort of 85 806 participants aged 40–79 years at recruitment from 12 southeastern states in the United States, recruited between 2002 and 2009 (9,10). Ascertainment of incident breast cancer cases among SCCS participants was conducted through annual linkage with state cancer registries in the SCCS catchment area. Data related to cancer diagnosis and treatment, including stage of diagnosis, first-course treatment, and tumor estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status were obtained from the state cancer registries.

Gene expression profiling and data processing

SCCS data

Gene expression levels in SCCS samples were derived from RNA-Seq of total RNA isolated from formalin-fixed paraffin-embedded (FFPE) breast cancer tissues. Total RNA was extracted and purified using Qiagen's miRNeasy FFPE Kit. The quantity and quality of the RNA samples extracted from tumor tissue FFPE sections were evaluated using Nanodrop (E260, E260/E280 ratio, spectrum 220–320 nm) and were separated on an Agilent BioAnalyzer. Both Ribo-Zero and RNase H were used for rRNA depletion. Illumina TruSeq RNA sample Prep Kit v2 was used to prepare a sequencing library, and HiSeq 2000 was used for sequencing. Each sample was sequenced pair-ended with a read length of 100 bp. A minimum of 10M reads was obtained for each sample. RNA-Seq data were processed following the mRNA analysis pipeline of TCGA GDC (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/). A two-pass method of STAR (11) was used for raw data alignment to the human reference genome (hg38). RNA expression levels were determined from aligned BAM files using HT-Seq-Count (12). The GENCODE v22 was used for coding gene and noncoding RNAs annotation in the human genome (13). Gene expression levels were measured using fragments per kilobase of transcripts per million mapped reads (FPKM), then $\log_2$-transformed after excluding genes that were not expressed in over half of the samples (median FPKM = 0). Quantile normalization was performed to standardize the expression level to the same scale. Probabilistic estimation of expression residuals factors was calculated to correct for batch effects and other potential experimental confounders (14). In this study, we only investigated protein-coding genes and lincRNAs.

The molecular intrinsic subtype of breast cancer (luminal A, luminal B, basal-like, HER2-enriched, and normal-like) was determined following the strategy of Giovanni's work using the PAM50 classifier (15,16) and implemented using R package *genefu* (17).

TCGA data

Breast cancer data from TCGA, which are publicly available from the NCI GDC data portal (https://portal.gdc.cancer.gov/), was used in this study as an independent validation set. We obtained the most updated clinical information from the TCGA Pan-Cancer Clinical Data Resource (18). HTSeq-FPKM data for TCGA breast cancer samples were downloaded as gene expression profiles and for further analysis, by using R package *TCGAbiolinks* (12,19). Principal component analyses were conducted using EIGENSTRAT bases on genotype data from TCGA samples (20). Using the 1000 Genomes Project data as reference (21), we used the first and second principal components of TCGA samples to determine race and kept those AA and EA samples for subsequent analyses.

TCGA RNA-Seq gene expression data were processed following the same transformation, normalization, and subtyping steps applied to the SCCS data. We used the TCGA breast cancer data as an external validation set for our race-differentiated gene signature identified from the SCCS.

Statistical analyses

All statistical analyses were performed in R 3.6.1. Differences in demographic and clinicopathological characteristics between AA and EA patients were assessed with *t*-test or nonparametric Wilcoxon rank-sum test for continuous variables. Binary and categorical data were analyzed with chi-square test. Univariable overall survival analyses were conducted using Kaplan–Meier estimates and log-rank tests. The Cox proportional hazards were used for multivariable modeling.

Gene signature development and validation

Using data from the SCCS, we identified genes that were differentially expressed between AA and EA breast cancer patients using linear regression models, adjusting for age at diagnosis, PAM50 subtypes, tumor, nodes, and metastases stages, probabilistic estimation of expression residuals factors, and batch effect. A false discovery rate (FDR) of 0.01 was employed to account for multiple testing. Penalized logistic regression with ridge regularization was used to reduce the prediction variance and overcome the potential problem of multicollinearity by shrinking the model coefficients when building the race-differentiated gene expression signature (22). Cross-validation was used for the tuning parameter selection.

To externally validate this gene signature derived from the SCCS, we used data from TCGA. The identified gene signature was fitted to the TCGA gene expression data to predict patient race (AA versus EA). Ridge regression was used instead of lasso or elastic net because the purpose was to validate the gene signature post model finalization; therefore no further variable selection should be performed at this step. The gene signature was then evaluated in terms of both discrimination and calibration. The C-statistic was calculated to assess the predictive accuracy of the entire signature in discriminating AA and EA breast cancer in this external validation set. Calibration of the gene signature was assessed graphically in a calibration graph by plotting the observed outcomes against the predicted probabilities.

Survival analyses

The race-differentiated genes identified above were assessed for their associations with breast cancer survival in the SCCS. Allowing for retention of correlated genes, we used elastic net penalized Cox regression to identify a subset of the race-differentiated genes that were prognostic of breast cancer survival, adjusting for clinical variables, including age at diagnosis, stage, race, ER, PR, and HER2. Variable selection was achieved through elastic net-penalized partial likelihood, where both mixing and overall shrinkage tuning parameters are simultaneously cross-validated.

Results

Study samples

The demographic and clinical characteristics are summarized separately for SCCS and TCGA patients in Table 1. A total of 415 SCCS female participants (260 AA and 155 EA) with breast cancer were included in the study. The median follow-up period in the SCCS was 132 months (range, 7–177 months). AA cases

Table 1. Clinicopathological characteristics of breast cancer patients according to race groups in the SCCS and TCGA

Variable	SCCS (*n* = 415)			TCGA (*n* = 1018)		
Race	AA	EA	*P* value	AA	EA	*P* value
No. of subjects (%)	260 (62.7)	155 (37.3)		180 (17.7)	838 (82.3)	
Age at diagnosis, mean (SD)	58.5 (9.0)	60.6 (9.4)	0.029	56.3 (13.5)	59.2 (13.1)	0.007
PAM50 subtype, no. (%)						
Luminal A	112 (43.1)	78 (50.3)	0.474	48 (26.7)	410 (48.9)	<0.001
Luminal B	36 (13.8)	24 (15.5)		34 (18.9)	180 (21.5)	
Basal-like	67 (25.8)	32 (20.6)		67 (37.2)	120 (14.3)	
HER2-enriched	29 (11.2)	12 (7.7)		22 (12.2)	75 (8.9)	
Normal-like	16 (6.2)	9 (5.8)		9 (5.0)	53 (6.3)	
TNBC, no. (%)						
Yes	47 (18.1)	21 (13.5)	0.285	33 (18.3)	73 (8.7)	<0.001
No[a]	213 (81.9)	134 (86.5)		147 (81.7)	765 (91.3)	
ER, no. (%)						
Negative	85 (33.3)	36 (24.8)	0.095	69 (39.2)	148 (18.6)	<0.001
Positive	170 (66.7)	109 (75.2)		107 (60.8)	648 (81.4)	
Undetermined[b]	5	10		4	42	
PR, no. (%)						
Negative	110 (43.7)	52 (36.1)	0.173	88 (50)	229 (28.9)	<0.001
Positive	142 (56.3)	92 (63.9)		88 (50)	564 (71.1)	
Undetermined[b]	8	11		4	45	
HER2, no. (%)						
Negative	165 (85.5)	97 (84.3)	0.914	74 (83.1)	450 (77.4)	0.378
Positive	28 (14.5)	18 (15.7)		16 (16.9)	131 (22.6)	
Undetermined[b]	67	40		91	257	
AJCC stage, no. (%)						
Stage I	74 (36.1)	65 (51.2)	0.02	32 (18.2)	145 (17.7)	0.614
Stage II	83 (40.5)	46 (36.2)		106 (60.2)	465 (56.6)	
Stage III	38 (18.5)	14 (11.0)		34 (19.3)	196 (23.9)	
Stage IV	10 (4.9)	2 (1.6)		4 (2.3)	15 (1.8)	
Stage X or unknown	55	28		4	19	

[a] All participants with undetermined for any or all the three components were included in the ‘No’ group.
[b] Undetermined or not evaluated—not included in statistical tests to evaluate racial differences.

were diagnosed at a younger age than EA cases in both the SCCS (*P* = 0.029) and TCGA (*P* = 0.007). Using the PAM50 classifier, more AA patients were classified as basal-like and HER2-enriched breast cancer subtypes than EA patients (basal-like: 25.8% versus 20.6% in the SCCS, 37.2% versus 14.3% in TCGA; HER2-enriched: 11.2% versus 7.7% in the SCCS, 12.2% versus 8.9% in TCGA). Fewer AA patients were classified as luminal A and luminal B breast cancer subtypes than EA patients (luminal A: 43.1% versus 50.3% [*P* = 0.154] in the SCCS, 26.7% versus 48.9% [*P* < 0.001] in TCGA; luminal B: 13.8% versus 15.5% [*P* = 0.638] in the SCCS, 18.9% versus 21.5% [*P* = 0.423] in TCGA). More AA patients were classified as triple negative breast cancer (TNBC) subtype than EA patients (18.1% versus 13.5% [*P* = 0.206] in the SCCS; 18.3% versus 8.7% [*P* = 0.0016] in TCGA). These differences were statistically significant in TCGA but not in the SCCS.

Race-differentiated gene signature in breast cancer tumor tissues

We identified 19 065 genes (16 586 protein-coding and 2479 lincRNAs) that were expressed in over half of the SCCS samples. Of these, 2001 (10.5%) were differentially expressed in EA and AA at a nominal *P* value < 0.05, among which 59 genes (54 protein-coding genes and 5 lincRNAs) reached an FDR-adjusted *P* value < 0.01 (Table 2). Of these 59 genes, 31 genes expressed significantly higher in AA than EA women, while the remaining 28 genes expressed significantly higher in EA than AA women. The top three differentially expressed genes which expressed higher in AA were TAS2R43 (OMIM 612668), NUTM2F (no OMIM ID), and PWP2 (OMIM 601475), while the top three differentially expressed genes which expressed higher in EA were PM20D1(OMIM 617124), BIRC7 (OMIM 605737), and LOC102724159 (ENSG00000275464, no OMIM). Hierarchical clustering of this gene set shows a good performance of the enrichment for both race and different subtypes (Figure 1). All 59 of these genes were included in constructing the race-differentiated gene signature. Penalized logistic regression model with ridge regularization was implemented to shrink the coefficients in order to reduce the variance in prediction.

External validation in TCGA

We validated our 59-gene signature by fitting the model to the independent TCGA gene expression data to predict patient race (AA versus EA). The gene signature was then evaluated in terms of both discrimination and calibration. The concordance statistic (C-statistic, equal to the area under the receiver operating characteristic curve, AUC) was computed to evaluate overall prediction accuracy. At the external validation in TCGA breast cancer data, the C-statistic for the 59-gene signature was 0.81, indicating high discriminative ability of the gene signature to distinguish between AA and EA breast cancer patients. Calibration is the degree of correspondence between the estimated probability produced by the gene signature and the actual observed probability. The calibration curve of the gene signature was constructed with TCGA data and shown in Supplementary Figure 1,

Table 2. The set of 59 genes found to be significantly differential expressed in the SCCS at FDR-adjusted *P* value < 0.01

ENSEMBL ID	Gene symbol	P value (SCCS)	FDR (SCCS)	P value (TCGA)
ENSG00000255374	TAS2R43	<0.001	<0.001	<0.001
ENSG00000130950	NUTM2F	<0.001	<0.001	<0.001
ENSG00000241945	PWP2	<0.001	<0.001	<0.001
ENSG00000259604	Lnc-ALDH1A3-1	<0.001	0.005	<0.001
ENSG00000160221	GATD3A	<0.001	<0.001	<0.001
ENSG00000176681	LRRC37A	<0.001	<0.001	<0.001
ENSG00000142178	SIK1	<0.001	0.001	<0.001
ENSG00000134184	GSTM1	<0.001	<0.001	<0.001
ENSG00000261701	HPR	<0.001	0.008	<0.001
ENSG00000158488	CD1E	<0.001	<0.001	<0.001
ENSG00000196436	NPIPB15	<0.001	<0.001	<0.001
ENSG00000173678	SPDYE2B	<0.001	<0.001	<0.001
ENSG00000224635	AL391095.1	<0.001	<0.001	<0.001
ENSG00000106113	CRHR2	<0.001	0.004	<0.001
ENSG00000105499	PLA2G4C	<0.001	<0.001	<0.001
ENSG00000204147	ASAH2B	<0.001	0.002	<0.001
ENSG00000197822	OCLN	<0.001	0.008	<0.001
ENSG00000197728	RPS26	<0.001	<0.001	<0.001
ENSG00000168255	POLR2J3	<0.001	<0.001	<0.001
ENSG00000203666	EFCAB2	<0.001	0.003	<0.001
ENSG00000106686	SPATA6L	<0.001	<0.001	<0.001
ENSG00000152689	RASGRP3	<0.001	0.008	<0.001
ENSG00000178295	GEN1	<0.001	<0.001	<0.001
ENSG00000215908	CROCCP2	<0.001	<0.001	<0.001
ENSG00000154237	LRRK1	<0.001	0.001	<0.001
ENSG00000026950	BTN3A1	<0.001	0.007	<0.001
ENSG00000149292	TTC12	<0.001	0.004	<0.001
ENSG00000198060	MARCH5	<0.001	0.004	<0.001
ENSG00000154096	THY1	<0.001	<0.001	<0.001
ENSG00000111671	SPSB2	<0.001	0.008	<0.001
ENSG00000135723	FHOD1	<0.001	0.004	<0.001
ENSG00000150477	KIAA1328	<0.001	0.009	<0.001
ENSG00000231389	HLA-DPA1	<0.001	0.008	<0.001
ENSG00000131669	NINJ1	<0.001	0.009	<0.001
ENSG00000181481	RNF135	<0.001	0.009	<0.001
ENSG00000203747	FCGR3A	<0.001	0.002	<0.001
ENSG00000144026	ZNF514	<0.001	0.001	<0.001
ENSG00000186468	RPS23	<0.001	<0.001	<0.001
ENSG00000175745	NR2F1	<0.001	0.009	<0.001
ENSG00000171503	ETFDH	<0.001	0.002	<0.001
ENSG00000068654	POLR1A	<0.001	<0.001	<0.001
ENSG00000185684	EP400P1	<0.001	<0.001	<0.001
ENSG00000279457	WASH9P	<0.001	0.004	<0.001
ENSG00000280670	CCDC163	<0.001	0.008	<0.001
ENSG00000149089	APIP	<0.001	<0.001	<0.001
ENSG00000273136	NBPF26	<0.001	<0.001	<0.001
ENSG00000140263	SORD	<0.001	<0.001	<0.001
ENSG00000277203	F8A1	<0.001	<0.001	<0.001
ENSG00000223865	HLA-DPB1	<0.001	<0.001	<0.001
ENSG00000280071	GATD3B	<0.001	<0.001	<0.001
ENSG00000019169	MARCO	<0.001	<0.001	<0.001
ENSG00000262202	Lnc-GRAP-3	<0.001	<0.001	<0.001
ENSG00000132207	SLX1A	<0.001	<0.001	<0.001
ENSG00000215374	FAM66B	<0.001	0.001	<0.001
ENSG00000016402	IL20RA	<0.001	<0.001	<0.001
ENSG00000170160	CCDC144A	<0.001	0.005	<0.001
ENSG00000275464	LOC102724159	<0.001	<0.001	<0.001
ENSG00000101197	BIRC7	<0.001	0.008	<0.001
ENSG00000162877	PM20D1	<0.001	<0.001	<0.001

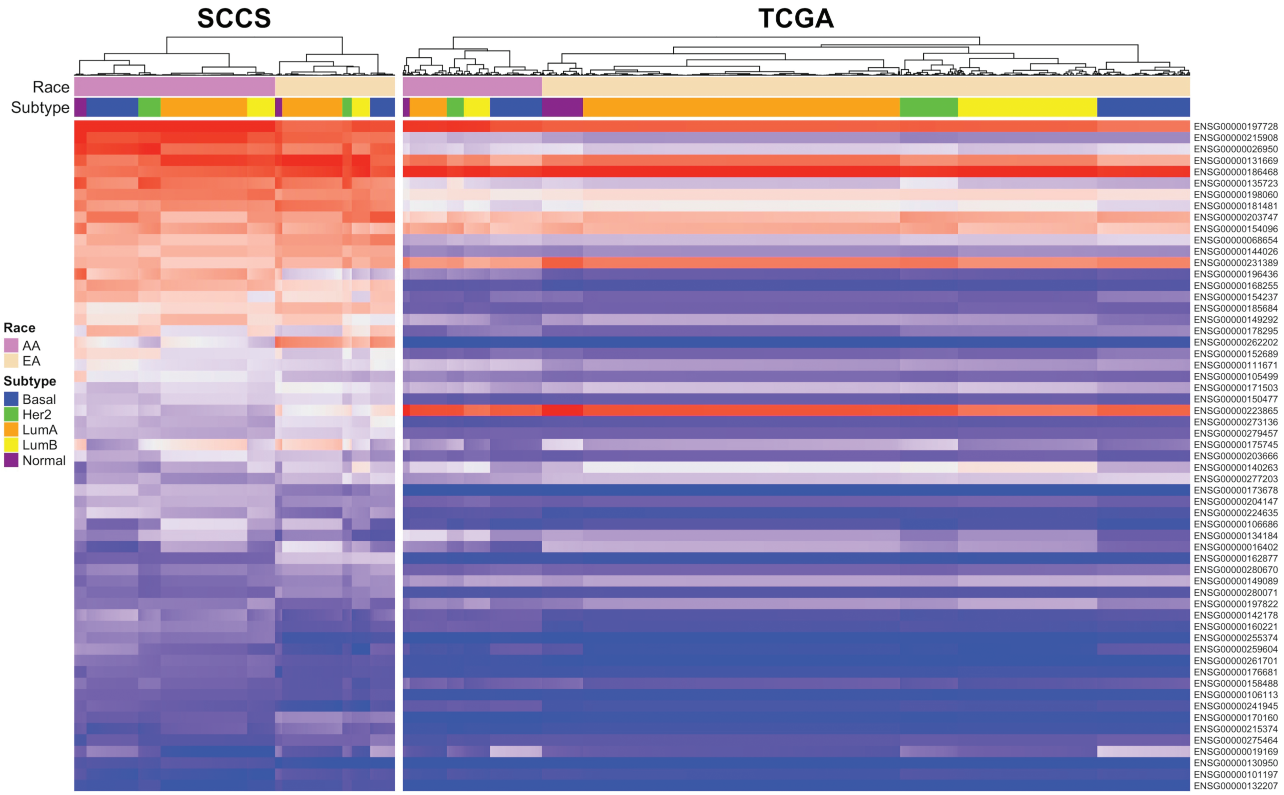

Figure 1. Hierarchical clustering of 59 genes significantly differentially expressed between AA and EA patients after adjusting for age at diagnosis, PAM50 subtypes, tumor, nodes, and metastases stages, probabilistic estimation of expression residuals factors and batch effects for both the SCCS and TCGA.

available at *Carcinogenesis* Online. The Brier score measures disagreement between the observed outcome and a prediction and is the mean squared error ranging from 0 to 1, where the lower Brier score indicates the better calibration of the predictions. The Brier score of our 59-gene signature was 0.064, indicating high reliability and prediction accuracy of the gene signature to distinguish between AA and EA breast cancer patients.

Survival analysis

Using the Cox proportional hazards regression model, adjusting for age at diagnosis, subtype, ER, PR, HER2, and stage, we investigated the association of each gene from our 59 race-differentiated genes with overall survival (Supplementary Table 1, available at *Carcinogenesis* Online) among both AA and EA women. Using a *P* value of 0.10 as the tentative threshold for a significant association in the SCCS, we found that 10 of the 59 genes were associated with overall survival in AA but not in EA, while 7 genes were associated with overall survival in EA but not in AA. None of these genes were significantly associated with breast cancer survival after adjusting for multiple comparisons.

Race-differentiated gene signature by breast cancer subtypes

We investigated differentially expressed genes between AA and EA patients by three different breast cancer subtypes: (i) basal-like breast cancer subtype; (ii) luminal breast cancer subtype; and (iii) TNBC subtype. The analysis strategy was the same as the overall breast cancer patients. Among those breast cancer patients with both basal-like and TNBC subtypes, no gene reached an FDR-adjusted *P* value < 0.05. Within the patients with luminal breast cancer subtype, 1847 (9.7%) were differentially expressed in EA and AA at a nominal *P* value < 0.05, among which 40 genes (36 protein-coding genes and 4 lincRNAs) reached an FDR-adjusted *P* value < 0.01 (Supplementary Table 2, available at *Carcinogenesis* Online). Twenty-seven of these 40 genes were in the overall 59 race-differentiated genes. We built a luminal-subtype-specific race-differentiated gene signature with these 40 genes by penalized logistic regression model with ridge regularization and externally validated with TCGA data. At the external validation in TCGA luminal breast cancer data, the C-statistic for the 40-gene signature was 0.80 and the Brier score was 0.21.

Discussion

In this study, we investigated differentially expressed genes between AA and EA breast cancer patients by using RNA-Seq data from the SCCS and using TCGA data for validation. For the first time, a race-differentiated gene signature was constructed and validated externally to predict breast cancer patient race. Moreover, we performed survival analyses to explore the association between gene expression and overall survival.

In total, we identified 59 genes that showed significant differences in their expression levels between AA and EA breast cancer at an FDR-adjusted *P* value < 0.01. The results described in our study were quite consistent with previous studies. For example, 19 of 59 differentially expressed genes identified in our study were also differentially expressed at an FDR-adjusted *P* value of < 0.05 in Huo's study (4).

Within the 59 race-differentiated genes, we found that several genes play a critical role in the immune system, such as IL20RA, MARCO, BTN3A1, HLA-DPA1 and HLA-DPB1. Some genes were related to the prognosis of breast cancer, such as SORD and FAM3A. SORD is a key enzyme in the polyol pathway and plays an important role in the development of diabetic complications (23). SORD is usually used as the measurement of epithelial-to-mesenchymal transition suppression. Polyol pathways represent a molecular link between glucose metabolism and cancer differentiation. The overexpression of SORD may cause poor prognoses of cancer (24). FAM3A belongs to a family of cytokines containing four genes (FAM3A, FAM3B, FAM3C and FAM3D) that are mainly expressed in highly proliferative tissue, and it has been suggested that it plays a core role in cell proliferation, and functions as an activator of the ERK1/2 and p38MAPK signaling pathways (25,26). Interestingly, PWP2 which was more highly expressed in AA and LOC102724159 which was more highly expressed in EA are paralogous genes. Among the genes that were strongly associated with overall survival in both the SCCS and TCGA data, SPSB2 regulated protein degradation by acting as adaptors for ubiquitin ligases (27).

Our study focused on the discovery of the biological basis behind the race/ethnic difference between AA and EA breast cancer patients instead of the discovery of specific genes and as the cumulative effect of multiple genes rather than a single gene effect determines cancer phenotypes, a combination of race-differentiated genes may simultaneously contribute to the biological difference of racial disparity in breast cancer. We performed a Gene Ontology (GO) biological processes enrichment analysis with the 59 race-differentiated genes (Supplementary Table 3, available at *Carcinogenesis* Online) using WebGestalt (28). Forty-three GO terms were enriched at nominal *P* value < 0.05 and most of these GO terms focus on endonuclease activity, endodeoxyribonuclease activity and antigen binding (Supplementary Figure 2, available at *Carcinogenesis* Online). Endonuclease activity may affect the base excision repair (BER) pathway (29), which was reported different between AA and EA (30).

There are two limitations in the present study. First, our RNA-Seq data in the SCCS were generated from FFPE samples, and thus, gene expression in some genes, particularly those with low expression levels, may not be reliably measured due to the processing and storage of FFPE samples (31,32). The second limitation is that only overall survival data were collected in the SCCS.

Conclusion

In this study, we analyzed racial differences in gene expression using RNA-Seq data generated from breast cancer samples included in the SCCS. We identified 59 that were differentially expressed in AA and EA breast cancer samples included in the SCCS. A gene signature constructed using these genes were externally validated using TCGA data yielding a C-statistic of 0.81 and the Brier score of 0.064. These findings provide insight into the biological differences in tumors and the survival disparity between AA and EA breast cancer patients.

Supplementary material

Supplementary data are available at *Carcinogenesis* online.

Supplementary Figure 1. Calibration plot of the 59-gene signature model with TCGA data to diagnose lack of fit. A calibration curve describes the relationship between predicted values (*x*-axis) and the truth (*y*-axis). The solid curve is for the fitted logistic regression model, and the dotted curve is its loess smoother. The diagonal line represents the line of perfect calibration. If the smoother lies close to the diagonal, the model is well calibrated. The smoother shows a slight deviation from the diagonal line. The Brier score of the gene signature was 0.064, indicating high reliability and prediction accuracy of the gene signature to distinguish between AA and EA breast cancer patients.

Supplementary Figure 2. Bar chart for enrichment ratios of 43 GO terms which were enriched at nominal *P*-value < 0.05 from 59 race-differentiated genes.

Supplementary Table 1. Association of selected genes with overall survival among breast cancer patients

Supplementary Table 2. The set of 40 genes found to be significantly differential expressed between AA and EA in the Luminal subtype breast cancer patients of SCCS at an FDR-adjusted *P* value < 0.01.

Supplementary Table 3. Gene Ontology biological processes enrichment results with the 59 race-differentiated genes

Funding

Research reported in this publication was supported in part by grants from the National Institutes of Health (R01CA202981 and U01CA202929). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SCCS data collection was performed by the Survey and Biospecimen Shared Resource which is supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA68485). The study used resources at the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, which is supported in part by the National Institutes of Health S10 Shared Instrumentation (1S10OD023680).

Acknowledgement

The authors thank Marshal Younger for assistance with editing and manuscript preparation.

Conflict of Interest Statement: None declared.

References

1. Bray, F. *et al.* (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.*, 68, 394–424.
2. Howlader, N.N.A. *et al.* (eds) (2019) *SEER Cancer Statistics Review, 1975–2016*. National Cancer Institute, Bethesda, MD. https://seercancergov/csr/1975_2016/ (October 2019, date last accessed).
3. Gupta, V. *et al.* (2018) Racial disparity in breast cancer: can it be mattered for prognosis and therapy. *J. Cell Commun. Signal.*, 12, 119–132.
4. Huo, D. *et al.* (2017) Comparison of breast cancer molecular features and survival by African and European ancestry in the Cancer Genome Atlas. *JAMA Oncol.*, 3, 1654–1662.
5. Martin, D.N. *et al.* (2009) Differences in the tumor microenvironment between African-American and European-American breast cancer patients. *PLoS One*, 4, e4531.
6. Field, L.A. *et al.* (2012) Identification of differentially expressed genes in breast tumors from African American compared with Caucasian women. *Cancer*, 118, 1334–1344.
7. Parada, H. Jr *et al.* (2017) Race-associated biological differences among luminal A and basal-like breast cancers in the Carolina Breast Cancer Study. *Breast Cancer Res.*, 19, 131.
8. D'Arcy, M. *et al.* (2015) Race-associated biological differences among luminal A breast tumors. *Breast Cancer Res. Treat.*, 152, 437–448.
9. Signorello, L.B. *et al.* (2010) The Southern Community Cohort Study: investigating health disparities. *J. Health Care Poor Underserved*, 21(1 suppl.), 26–37.
10. Signorello, L.B. *et al.* (2005) Southern Community Cohort Study: establishing a cohort to investigate health disparities. *J. Natl. Med. Assoc.*, 97, 972–979.

11. Dobin, A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15–21.
12. Anders, S. *et al.* (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 166–169.
13. Harrow, J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22, 1760–1774.
14. Stegle, O. *et al.* (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, 6, e1000770.
15. Parker, J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27, 1160–1167.
16. Ciriello, G. *et al.*; TCGA Research Network. (2015) Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163, 506–519.
17. Gendoo, D.M. *et al.* (2016) Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*, 32, 1097–1099.
18. Liu, J. *et al.*; Cancer Genome Atlas Research Network. (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173, 400.e11–416.e11.
19. Colaprico, A. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, 44, e71.
20. Price, A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38, 904–909.
21. Auton, A. *et al.*; 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, 526, 68–74.
22. Jain, R.K. (1985) Ridge regression and its application to medical data. *Comput. Biomed. Res.*, 18, 363–368.
23. Carr, I.M. *et al.* (1995) Molecular genetic analysis of the human sorbitol dehydrogenase gene. *Mamm. Genome*, 6, 645–652.
24. Schwab, A. *et al.* (2018) Polyol pathway links glucose metabolism to the aggressiveness of cancer cells. *Cancer Res.*, 78, 1604–1618.
25. Peng, X. *et al.* (2016) Identification of FAM3D as a new endogenous chemotaxis agonist for the formyl peptide receptors. *J. Cell Sci.*, 129, 1831–1842.
26. Zhu, Y. *et al.* (2002) Cloning, expression, and initial characterization of a novel cytokine-like gene family. *Genomics*, 80, 144–150.
27. Kuang, Z. *et al.* (2010) The SPRY domain-containing SOCS box protein SPSB2 targets iNOS for proteasomal degradation. *J. Cell Biol.*, 190, 129–141.
28. Liao, Y. *et al.* (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*, 47, W199–W205.
29. Kim, Y.J. *et al.* (2012) Overview of base excision repair biochemistry. *Curr. Mol. Pharmacol.*, 5, 3–13.
30. Gao, R. *et al.* (2008) Ethnic disparities in Americans of European descent versus Americans of African descent related to polymorphic ERCC1, ERCC2, XRCC1, and PARP1. *Mol. Cancer Ther.*, 7, 1246–1250.
31. Groelz, D. *et al.* (2013) Non-formalin fixative versus formalin-fixed tissue: a comparison of histology and RNA quality. *Exp. Mol. Pathol.*, 94, 188–194.
32. von Ahlfen, S. *et al.* (2007) Determinants of RNA quality from FFPE samples. *PLoS One*, 2, e1261.