Breast Heterogeneity: Obstacles to Developing Universal Biomarkers of Breast Cancer Initiation and Progression

Rebecca C Dirks, MD, Heather N Burney, MS, Manjushree Anjanappa, MS,
George E Sandusky, DVM, PhD, Yangyang Hao, PhD, Yunlong Liu, PhD, Max C Schmidt, MD, PhD,
Harikrishna Nakshatri, BVSc, PhD

BACKGROUND: Predicting outcomes and response to therapy through biomarkers is a major challenge in cancer research. In previous studies, we suggested that inappropriate “normal” tissue samples used for comparison with tumors, inter-individual heterogeneity in gene expression, and genetic ancestry all influence biomarker expression in tumors. The aim of this study was to investigate these factors in breast cancer using breast tissues from healthy women and normal tissue adjacent to tumor (NAT) with matrix metalloproteinase 7 (MMP7) as a candidate biomarker.

STUDY DESIGN: RNA sequencing was performed on primary luminal progenitor cells from healthy breast, NATs, and tumors to identify transcriptomes enriched in NATs and breast cancer. Expression of select genes was validated via quantitative reverse transcription polymerase chain reaction of RNA and via immunohistochemistry of a tissue microarray of normal, NAT, and tumor samples of different genetic ancestry.

RESULTS: Twenty-six genes were significantly overexpressed in NATs and tumors compared with healthy controls at messenger RNA level and formed a para-inflammatory network. MMP7 had the greatest expression in tumor cells, with upregulation confirmed by quantitative reverse transcription polymerase chain reaction. Tumor-enriched but not NAT-enriched expression of MMP7 compared with healthy controls was reproduced at protein levels. When stratified by genetic ancestry, tumor-specific increase of MMP7 reached statistical significance in women of European ancestry.

CONCLUSIONS: Transcriptome differences across healthy, NAT, and tumor tissue in breast cancer demonstrate an active para-inflammatory network in NATs and indicate unsuitability of NATs as “normal controls” in biomarker discovery. The discordance between transcriptomic and proteomic MMP7 expression in NATs and the influence of genetic ancestry on its protein expression highlight the complexity in developing universally acceptable biomarkers of breast cancer and the importance of genetic ancestry in biomarker development. (J Am Coll Surg 2020;231:85—97. © 2020 by the American College of Surgeons. Published by Elsevier Inc. All rights reserved.)

Disclosure Information: Nothing to disclose.

Support: Dr Dirks is supported by the Shirley Dawn Wyss Breast Cancer Fellowship at the Indiana University Foundation. Dr Nakshatri received support for this work from Susan G Komen for the Cure (SAC110025), Department of Defense (W81XWH-15-1-0707), and 100 Voices of Hope. Susan G Komen for the Cure, Breast Cancer Research Foundation, Vera Bradley Foundation for Breast Cancer Research, and the Catherine Peachy Foundation provide funding support to the Komen Normal Tissue Bank.

Presented at the Western Surgical Association 127th Scientific Session, Las Vegas, NV, November 2019.

Received January 6, 2020; Revised March 17, 2020; Accepted March 18, 2020.

From the Departments of Surgery (Dirks, Anjanappa, Schmidt, Nakshatri), Biostatistics (Burney), Pathology and Laboratory Medicine (Sandusky), Medical and Molecular Genetics (Hao, Liu), Biochemistry and Molecular Biology (Nakshatri), Indiana University School of Medicine, and Roudebush Veterans Affairs Medical Center (Nakshatri), Indianapolis, IN.

Correspondence address: Harikrishna Nakshatri, BVSc, PhD, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, C218C, 980 West Walnut St, Indianapolis, IN 46202. email: hnakshat@iupui.edu

https://doi.org/10.1016/j.jamcollsurg.2020.03.035
ISSN 1072-7515/20

Abbreviations and Acronyms

AA	=	African American
EA	=	European American
ER	=	estrogen receptor
IHC	=	immunohistochemistry
MMP7	=	matrix metalloproteinase 7
mRNA	=	messenger RNA
NAT	=	normal tissue adjacent to tumor
qRT-PCR	=	quantitative reverse transcription polymerase chain reaction

The search for new prognostic and predictive biomarkers of cancer remains ubiquitous. This area of research has stretched across cancers, including epithelial ovarian,[1] endometrial,[2] cervical,[3] and prostate cancer,[4] as well as many others within just the past few years.[5-8] Breast cancer treatment also has a burgeoning search for new biomarkers,[9] including metastatic spread,[10] therapy response,[11] and, more recently, adequate markers of immunotherapy response.[12,13] Although biomarkers associated with early breast cancer detection have exciting potential to improve patient outcomes, universal biomarkers remain deficient.[14] This study used para-tumoral breast tissue and genetic ancestry-considerate analysis to improve the search for biomarkers.

Tissues neighboring a tumor can hold a key to early cancer recognition. Nonmalignant, normal tissue adjacent to tumor (NAT) is pathologically benign yet is abnormal. For example, our group has recently demonstrated enrichment of ZEB1+ cells in the NATs of women of European ancestry,[15] and other groups have shown distinct DNA methylation patterns and epigenetic changes in genes related to stemness-associated signaling networks in NATs.[16] These abnormalities in NATs emphasize the need for true healthy tissues as "controls" in biomarker research and provide a resource for biomarker discovery.

Cancer-induced inflammation could be responsible for some of the epigenomic changes seen in NATs. Inflammation has been described as one of the hallmarks of cancer.[17] In para-inflammation, epithelial cells themselves express genes linked to inflammation and the immune system.[18] Although not malignant, reprogrammed epithelial cells can contribute to tumor initiation and progression[19] and para-inflammatory changes within NATs could serve as biomarkers. Earlier research from our group has illustrated genetic ancestry-dependent differences in such cancer-induced field defects in NATs.[15]

The goals of this study were to demonstrate that NAT tissue is abnormal and to use this abnormality as well as genetic ancestry-considerate analysis to improve biomarker discovery. We hypothesized that comparing transcriptomes from NATs to those of both healthy breasts and breast cancer samples would reveal para-inflammatory biomarkers of breast cancer, and that use of ancestry-considerate analysis would impact biomarker discovery. Transcriptomes were generated from purified luminal progenitors of healthy normal, NAT, and tumor tissue to limit the effects of differences in differentiation status between tissue types, given our previous study had shown remarkable inter-individual differences in stem-progenitor-mature/differentiated cell hierarchy,[20] more than 2,000 genes are differentially expressed between luminal progenitors and mature luminal cells,[21] and the majority of breast cancers are suggested to originate from luminal progenitors.[22,23] After obtaining results from transcriptomic data, we focused our attention on matrix metalloproteinase 7 (MMP7), a member of the MMP family of zinc-dependent endopeptidases, for further exploration at messenger RNA (mRNA) and protein levels.

METHODS

Primary cell lines and culture

Breast epithelial cells from de-identified healthy tissue containing core biopsies donated by healthy women were obtained from the Komen Tissue Bank at Indiana University. Samples with diverse genetic ancestry were sought for inclusion. De-identified tumor tissues and NATs were obtained from surgical cases at Indiana University after written consent, based on availability. All healthy, NAT, and tumor primary epithelial cells came from either fresh or cryopreserved breast tissues. Primary breast epithelial cells for RNA sequencing were propagated using a previously described epithelial cell reprogramming assay.[24] For validation of data by quantitative reverse transcription polymerase chain reaction (qRT-PCR), we used an improved method developed in the laboratory, which does not require irradiated mouse embryonic fibroblast feeder layer. Briefly, this method used culture dishes precoated with laminin-5-rich-conditioned media from 804G cell line and a growth media supplemented with inhibitors of Rho kinase, transforming growth factor-β, and bone morphogenetic protein signaling.[25]

RNA sequencing and Ingenuity Pathway Analysis

RNA sequencing was performed on primary cells from healthy, NAT, and tumor tissue. All primary cells were sorted by flow cytometry before sequencing to enrich for luminal progenitor cells. Combination CD49f+ and epithelial cell adhesion molecule-positive cells were selected to obtain this population.[26] This selection is necessary because of enormous inter-individual differences in stem-luminal progenitor-mature cell hierarchy of the normal breast impacting the number of progenitor cells at a given time.[20] Luminal progenitor cells were chosen in particular because the majority of breast cancer subtypes, including basal subtype, are suggested to originate from luminal progenitor cells.[26] CD49f/epithelial cell adhesion molecule-staining patterns of 3 samples from healthy breast, 3 paired NATs, and tumors and ductal carcinoma in situ and lobular carcinoma in situ of the same patient have been presented previously.[20] Representative staining patterns of a few samples gated for flow sorting are shown in Figure 1A. After assessing the concentration and quality of total RNA in samples, a complementary DNA library was created for sequencing. RNA sequencing was performed as described previously.[27] Differential expression analysis was accomplished using *edgeR* package and both p values and false discovery rate were computed. Ingenuity Pathway Analysis (Qiagen) was used to characterize the genomic changes found with differential expression analysis. Genes identified by RNA sequencing with $p < 0.001$ and false discovery rate < 0.05 were imported into Ingenuity Pathway Analysis and networks with associated diseases and functions were noted.

Quantitative reverse transcription polymerase chain reaction validation

Primary cells were cultured without earlier selection by flow cytometry, and RNA from exponentially growing cells was isolated with RNeasy Kit (74106; Qiagen). Complementary DNA from 2 μg of RNA for each sample was then created using the Bio-Rad iScript cDNA Synthesis Kit (170-8891). The Taqman universal PCR mix was used to perform qRT-PCR. For matched pairs of NAT and tumor cells, analysis of resultant qRT-PCR data was performed using the $\Delta\Delta C_T$ analysis described by Livak and Schmittgen.[28] β-Actin was the reference gene, and each tumor sample was normalized to its paired NAT sample to calculate relative fold change in MMP7 expression. Primers included ACTB (Hs01060665_g1) and MMP7 (Hs01042796_m1) (Applied Biosciences). For nonpaired samples, ΔC_T values were directly used in statistical analysis.

Immunohistochemistry

Immunohistochemistry staining evaluating the immunomarker MMP7 was performed on tissue microarrays containing cores of breast tissues donated by healthy women to the Komen Tissue Bank, NATs, and breast cancer from African-American (AA) and European-American (EA) women. This tissue microarrays has been described in detail previously.[15] All tissue collected was in compliance with an IRB-approved protocol, informed patient consent, and Health Insurance Portability and Accountability Act of 1996 protocol. A Clinical Laboratory Improvements Amendment-certified pathology core was used for immunohistochemistry and 3 blinded pathologists used light microscopy (Leica) to judge the intensity of MMP7 immunostain in each tissue core. Both positivity and H-score for each core were given. Any artifact, such as hemosiderin pigment, inflammation, surgical ink, and hemorrhage, was removed from the database and labeled on provided Excel (Microsoft) master database. Additionally, any cores with extensive physical damage (tears, folds) were excluded from analysis.

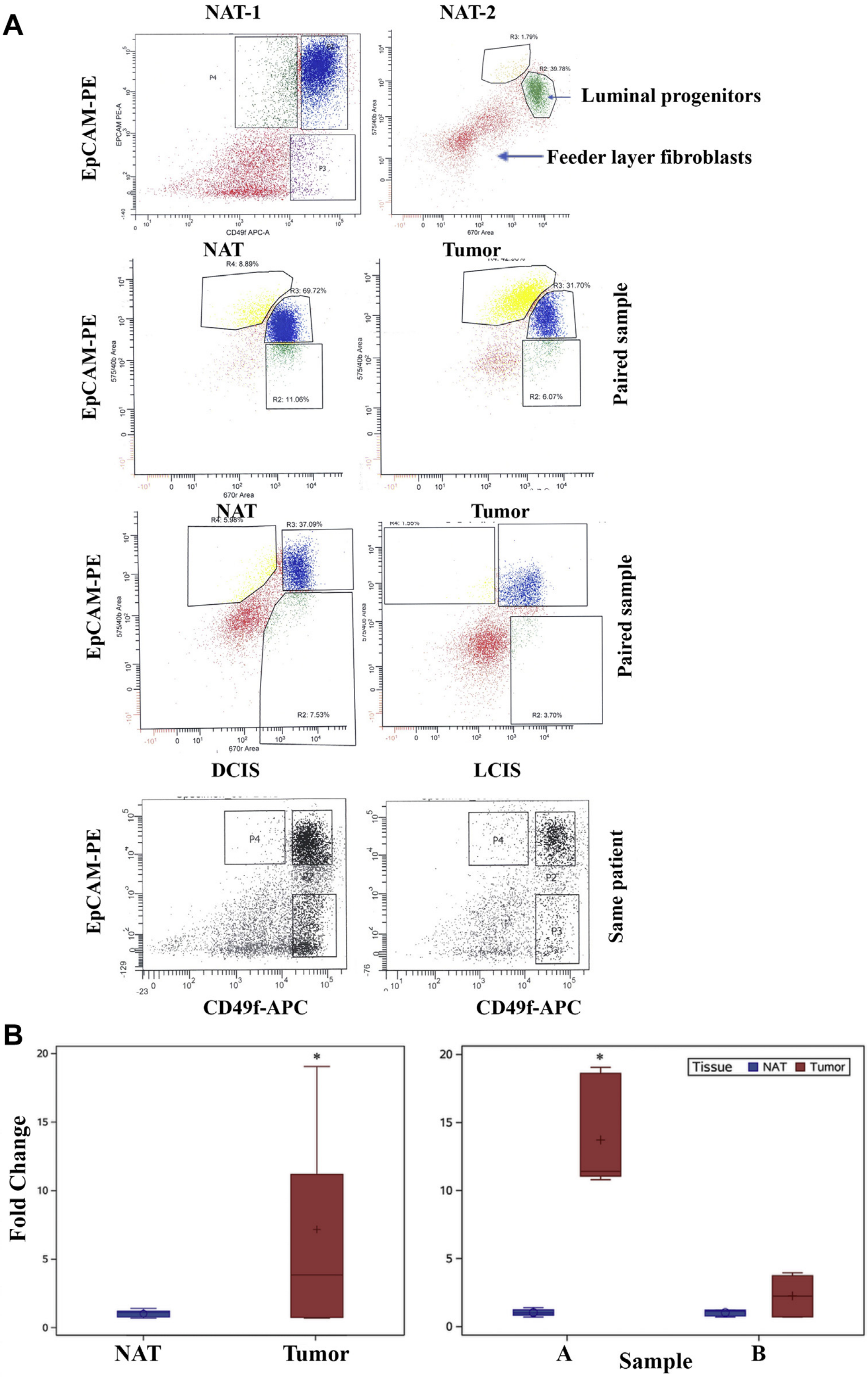
Statistical analysis

All experiments were conducted in at least 3 biologic replicates, except when limited by the replicative ability of primary cells. Instances where this was not possible are noted. Statistical analysis for differential expression analysis of RNA sequencing was performed in *edgeR*. Analysis of remaining data was performed on SAS University Edition, version 2018 (SAS Institute). Paired tumor and NAT samples were analyzed with Wilcoxon signed rank test, ordinal qRT-PCR data were analyzed with Wilcoxon rank sum (stage), remaining univariable analyses were performed with Kruskal-Wallis tests, and multivariable analyses were done with ANOVA. Univariable analysis was performed for nodal status, grade, and stage to avoid issues with collinearity. An α value of 0.05 was set for statistical significance ($p < 0.05$, false discovery rate < 0.05). MMP7-positivity and H-scores were analyzed by nonparametric Wilcoxon rank sum tests. All statistically significant differences are described in the text, while acknowledging that small sample size might have affected these analyses in some instances.

RESULTS

Inflammation-associated genes are upregulated in normal tissues adjacent to tumor and tumors compared with healthy breast epithelial cells

RNA sequencing was performed on primary cells from 12 breast tissue samples, including healthy, NAT, and tumor cells. There were 9 EA, 1 Hispanic, and 2 AA samples. EA samples included 2 matched NAT-tumor pairs. Hormone receptor status was known for tumor samples. Except for 1 NAT sample, all RNA-sequencing samples were from estrogen receptor α-positive (ER+) and progesterone receptor-positive breast cancer. None were human epidermal growth factor receptor 2-positive. Stage, nodal status, and grade were available for all tumor samples and


A
NAT-1
NAT-2
EpCAM-PE
P4
P3
EPCAM PE-A
CD49f APC-A
R3: 1.79%
R2: 39.78%
Luminal progenitors
Feeder layer fibroblasts
575/40b Area
670r Area
NAT
Tumor
R4: 8.89%
R3: 69.72%
R2: 11.06%
R3: 31.70%
R2: 6.07%
Paired sample
NAT
Tumor
R4: 5.98%
R3: 37.09%
R2: 7.53%
R2: 3.70%
Paired sample
DCIS
LCIS
P4
P2
P3
Same patient
CD49f-APC
CD49f-APC
B
Fold Change
20
15
10
5
0
*
NAT
Tumor
Tissue
NAT
Tumor
A
Sample
B

all but 1 NAT sample used for RNA sequencing. The percentage of luminal progenitor cells in tumors ranged from 40% to 87%. Although transcriptome analysis of only luminal progenitor cells from each group might have excluded several tumor cell-enriched transcripts, it also reduced the likelihood of identifying transcripts that are differentially expressed only due to variations in the differentiation status of healthy tissue, NATs, and tumors.

Differential expression analysis demonstrated a change in expression for 13,537 unique genes (eAppendix 1). When filtered for both significant false discovery rate (< 0.05) and significant p value (< 0.001), 26 unique genes remained. All were consistently overexpressed in NATs compared with healthy cells and further enriched in tumor cells (Table 1), and MMP7 had the greatest expression in tumor samples. Detailed expression levels for other genes in each of the samples are provided in Table 1. The subgroup of paired tumor vs NAT samples also showed upregulation in the tumor samples in the vast majority of these genes. When nonparametric statistical analysis was performed separately on nodal status, grade, and stage, these variables were not statistically associated with MMP7 expression.

When the 26 genes upregulated in NAT samples and further enriched in tumor samples were queried with Ingenuity Pathway Analysis, 2 networks were generated with 99% confidence that a similar network could not be created by random chance. These networks had 3 clusters of associated diseases and functions (Table 2). Immune and inflammatory diseases and functions were in both networks, and MMP7 was part of network 1.

Validating RNA-sequencing data of matrix metalloproteinase 7 by quantitative reverse transcription polymerase chain reaction

RNA sequencing results for MMP7 were validated in nonsorted cells via qRT-PCR. Given the limited replicative capacity of primary cells and low RNA yield, a limited number of genes could be validated. ITGAM and REN (Table 1) were initially chosen in addition to MMP7, as their expression is altered in breast cancer based on analyses of public databases.[29] Upregulation of these genes was not validated in preliminary PCR, so further use of limited primary cells was not pursued for these genes. Instead, MMP7 was chosen because it had the highest expression in tumor samples in our RNA-sequencing data, suggesting greater relevance as a breast cancer marker.

MMP7 qRT-PCR was performed on the 2 paired tumor-NAT samples from RNA sequencing (EA, ER+) in biologic quadruplicate. PCR demonstrated overexpression of MMP7 in the tumor samples, with median fold change of 3.8 ($p = 0.0274$) overall and median fold change of 2.2 ($p = 0.273$) and 11.4 ($p < 0.001$) when stratified by the individual donor (Fig. 1B), similar to RNA-sequencing data for the matched pairs.

Primary epithelial cells were then cultured from additional tumor, NAT, and healthy samples for qRT-PCR validation of MMP7. Except for 1 tumor sample collected in biologic duplicate due to limited replicative potential, RNA from the other samples was at least in biologic triplicates. In total, there were 9 tumor samples (5 EA, 3 AA, 1 Hispanic), 6 NAT samples (4 EA, 1 AA, 1 Hispanic), and 6 healthy samples (2 each EA, AA, and Hispanic). The healthy samples included 5 Komen Tissue Bank samples and 1 contralateral prophylactic breast tissue sample. When all samples were tested for MMP7 overexpression, there was a statistically significant difference in ΔC_T values across healthy, NAT, and tumor samples ($p = 0.0035$). In multivariable analysis, tissue type, genetic ancestry, and the interaction between these 2 variables (tissue by genetic ancestry interaction) all significantly influenced MMP7 level ($p = 0.014$, $p < 0.001$, and $p < 0.001$, respectively). This difference across both variables can be seen in Figure 2 where ΔC_T is inversely proportional to mRNA expression. While tumors of AA women, and to some extent EA women, expressed higher levels of MMP7 compared with NATs or healthy, tumors of Hispanic women contained lower levels of MMP7. These results were approached with caution, given the

Figure 1. Characterization of epithelial cells derived from healthy breast, normal tissue adjacent to tumor (NATs), and tumors. (A) CD49f/epithelial cell adhesion molecule staining pattern and gating of luminal progenitor cells for flow sorting of representative samples for RNA sequencing. Upper-right quadrant corresponds to luminal progenitor population. (B) Quantitative reverse transcription polymerase chain reaction validation of matrix metalloproteinase 7 (MMP7) expression in 2 paired NAT tumor samples. Relative fold change was calculated using $\Delta\Delta C_T$ analysis with β-actin reference gene and normalizing to NAT samples. Both samples were from breasts with estrogen receptor-positive tumors and women of European genetic ancestry. Left: The *MMP7* expression in tumor samples vs NAT samples for 2 matched pairs had a median fold change of 3.8 (*$p = 0.0274$, Wilcoxon signed rank test). Right: When stratified by individual, the *MMP7* expression had a median fold change of 11.4 for woman A (* $p \leq 0.001$, 2-way ANOVA) and a median fold change of 2.2 for woman B ($p > 0.05$, 2-way ANOVA).

Table 1. RNA Sequencing of Healthy Breast, Normal Adjacent to Tumor, and Tumor-Derived Luminal Progenitor Cells

	Mean counts per million			Log fold change			
Gene	Tumor (n = 5)	NAT (n = 3)	Healthy (n = 4)	NAT vs healthy	Tumor vs healthy	Tumor vs NAT	Paired tumor vs NAT
MMP7	40.949	4.164	0.801	2.271*	5.684*	3.404*	3.782*
PTPRJ	18.787	4.479	1.161	1.857*	4.087*	2.215*	2.842
STAB1	11.575	4.073	0.853	2.18*	3.840*	1.643	3.266
REN	10.956	3.072	1.173	1.333	3.539*	2.204*	3.163
GATM	8.064	1.834	0.683	1.354	3.665*	2.294*	3.415
CYBB	7.065	1.612	0.510	1.575	3.856*	2.262*	3.039
VTRNA1-2	6.767	3.361	1.302	1.263	2.847*	1.589	-2.745
EMR1	6.273	1.347	0.180	2.800*	5.191*	2.366*	3.089
ITGAM	6.056	1.670	0.178	3.119*	5.129*	1.990	2.609
PTPRC	3.110	0.874	0.250	1.711	3.694*	1.962	2.939
HCK	3.057	0.796	0.177	2.061*	4.153*	2.070	2.787
CP	2.415	0.460	0.013	4.932*	7.443*	2.521*	2.999
TNFRSF11B	2.071	0.452	0.146	1.526	3.821*	2.285*	3.060
TLR8	1.996	0.587	0.161	1.763	3.727*	1.945	2.700
SLC11A1	1.779	0.607	0.159	1.825	3.52*	1.676	2.633
MARCO	1.740	0.413	0.008	5.298*	7.502*	2.214*	2.714
TBXAS1	1.711	0.552	0.147	1.808	3.622*	1.800	2.821
C4orf26	1.688	0.207	0.029	2.675	5.824*	3.14*	3.868*
IRG1	1.340	0.085	0.000	5.687	9.608*	4.08*	4.468*
NEK5	1.281	0.130	0.012	3.130	6.547*	3.415*	4.1*
MMP12	0.936	0.169	0.037	2.026	4.617*	2.578*	2.748
CFTR	0.872	0.100	0.000	5.904*	8.979*	3.235*	3.565
ITGAD	0.784	0.148	0.028	2.209	4.727*	2.507*	3.221
CD207	0.630	0.079	0.000	5.571	8.684*	3.289*	2.775
CCL2	0.525	0.125	0.020	2.415	4.581*	2.164	2.677
BTK	0.501	0.116	0.012	2.969	5.172*	2.204	3.030

All fold changes p value < 0.001 and false discovery rate < 0.05.
*False discovery rate < 0.001.
NAT, normal tissue adjacent to tumor.

limited sample size, especially from Hispanic and AA women. Unfortunately, due to limited number of tumor samples from Hispanic women in our tissue repository, we could not pursue this observation further.

Tumor samples and NAT samples used for qRT-PCR also had known ER status. There were 7 ER-negative samples (4 NAT, 3 tumor) and 7 ER-positive samples (2 NAT, 5 tumor). All ER-positive samples were also progesterone receptor-positive, there were no human epidermal growth factor receptor 2-positive samples, and there was incomplete information on grade. When ER status, tissue type, and genetic ancestry were examined in multivariable analysis, ER status on its own was not statistically significant ($p = 0.1547$), but the interaction terms between both ER status and tissue type and between ER status and genetic ancestry were statistically significant (both, $p < 0.001$). Grade and nodal status were missing from many NAT and tumor samples, and NAT samples were largely missing the stage of the original tumor to which they were adjacent; all tumor samples had an associated stage. MMP7 ΔC_T value significantly correlated with stage, with stage IV samples having greater ΔC_T values, and smaller MMP7 expression than stage I or II ($p < 0.001$). There were no stage III tumor samples cultured for qRT-PCR.

Discordance between matrix metalloproteinase 7 transcripts and protein in normal breasts, normal tissues adjacent to tumor tissue, and tumors

A tissue microarray with 67 tumors, 23 NATs, and 33 healthy breasts undamaged cores was used for immunohistochemistry (IHC) with MMP7 antibody. MMP7 staining demonstrated a range of positivity, as seen in Figure 3. Staining was mainly noted in the tumor cell cytoplasm and in breast epithelial cells and ductal epithelial cells in the normal tissue samples. There was some staining of few lymphocytes, macrophages, fibroblasts, smooth muscle cells, and vascular endothelial cells. All cores from healthy EA women were positive, and 94% of the core from AA women were positive.

MMP7-positivity and H-scores were analyzed for the following parameters: tissue type, ER status, and genetic ancestry. In contrast to transcriptome data with luminal progenitor cells, we did not observe elevated MMP7 protein levels in NATs compared with healthy breast (Fig. 4). Surprisingly, NATs of AA women had lower levels of MMP7 compared with healthy counterparts ($p = 0.0126$), directly opposite the trend in mRNA in RNA sequencing. However, qRT-PCR and IHC data are compatible (Figs. 2 and 4) with lower expression in NATs compared with healthy tissues of AA women. Additionally, MMP7 H-score was still statistically increased in all tumor samples compared with NATs and healthy samples ($p = 0.0067$, $p = 0.0031$, respectively).

NATs and tumors were compared and stratified by genetic ancestry and ER status (Table 3). MMP7 H-score showed borderline elevated expression in tumors compared with NATs in AA women ($p = 0.0517$), with significantly elevated expression in EA women ($p = 0.0181$). Tumor and NAT differences were significant specifically in ER-positive samples of EA women and ER-negative samples of AA women. Comparisons between AA and EA samples, as stratified by tissue type and ER status, are shown in Table 4 and demonstrate similar expression in healthy tissue but significantly different expression in NATs and tumor tissues.

DISCUSSION

Radiologic techniques remain the mainstay of breast cancer detection, yet these methods can have high rates of false positives and negatives. Examination of the Dutch national registry, focused on a high-risk MRI screening program, revealed that 31% of breast cancers that had “negative” MRI 0 to 2 years before cancer detection had MRI-detectable cancers that were missed, and 34% of cases showed minimal signs.[30] Complementary molecular assays may improve earlier cancer detection. This study was initiated with a goal of using NATs for such assays, given earlier studies on molecular changes in NATs as biomarkers.[15]

Development of molecular biomarkers for early detection is exceedingly difficult because of enormous transcriptomic heterogeneity observed between healthy individuals due to enrichment of single nucleotide polymorphisms in gene regulatory regions.[31] To address some of this heterogeneity in our initial RNA-sequencing screen, we took a systematic approach of first purifying luminal progenitor cells of healthy breasts, NATs, and tumors to ensure that gene expression differences between the 3 tissue types were not due to inter-individual differences in stem-luminal progenitor-mature/differentiated cell hierarchy. This stringent criterion ultimately enabled us to select 26 genes upregulated in NATs and tumors compared with healthy breasts for additional analyses.

Although networks created from Ingenuity Pathway Analysis of the 26 genes involved diseases and functions intuitively associated with the breasts, such as lipid metabolism and connective tissue development and function (Table 2), these networks also repetitively included inflammatory and immunologic diseases and functions. These results support the concept of para-inflammation in tissues adjacent to tumors by demonstrating a consistent increase in inflammation-associated transcriptome in NATs compared with a more appropriate control (healthy samples from the Komen Tissue Bank). Because these transcriptome networks were generated using isolated cells and are less likely influenced by the microenvironment, these results indicate that NATs have undergone significant genomic changes to acquire para-inflammation[18] and NATs should not be used as “normal” controls in biomarker discovery.

Two members of the MMP family of zinc-dependent endopeptidases, MMP7 and MMP12, were notable members of Ingenuity Pathway Analysis inflammatory networks of NATs. This corroborates earlier literature concerning MMPs, inflammation, and cancer. As summarized in a 2017 review on MMPs by Alaseem and colleagues,[32] MMPs have an established connection with malignancy, including proliferation, invasion, angiogenesis and metastasis, and modulate inflammatory events to worsen pathologic conditions. Although MMP12 was also found to be upregulated in NATs and tumor samples (Table 1), MMP7 has previously been highlighted for its unique role in the MMP family as a signaling molecule and growth factor in addition to having the enzymatic activity shared by the rest of the family.[33] Additionally, MMP7 has prognostic potential,[34] and, as a secreted protein, there is also long-term potential to develop a plasma and serum-based ELISA to detect elevated MMP7 in cancer patients. Given these qualities and its notably greater upregulation, MMP7 was investigated further.

Genetic ancestry can influence the field effects found in NATs. For example, we have previously reported genetic ancestry might contribute to some of the differences in stem-luminal progenitor-mature cell hierarchy of the normal breasts.[20] In this study, genetic ancestry influenced MMP7 expression in both RNA and protein levels

Table 2. Ingenuity Pathway Analysis Networks and Associated Top Diseases and Functions

Network	Shared top disease and function	Included gene from RNA sequencing
1	Cellular movement, immune cell trafficking, inflammatory response, cellular function and maintenance	*ADGRE1, BTK, CCL2, CFTR, HCK, ITGAM, MMP7, MMP12, PTPRC, PTPRJ, SLC11A1, TLR8, TNFRSF11B*
2	Hematologic system development and function, inflammatory response, tissue morphology	*CP, CYBB, ITGAD, ODAPH, REN, STAB1, VTRNA1-2*
2	Connective tissue development and function, lipid metabolism, small molecule biochemistry	*CP, CYBB, ITGAD, ODAPH, REN, STAB1, VTRNA1-2*

(Tables 3 and 4 and Figs. 1, 2, and 4). Despite variations in the degree of significance, however, there was a consistent trend of higher MMP7 expression in tumors for both AA and EA women, and MMP7 H-score was statistically increased in all tumor samples compared with NATs and to healthy samples ($p = 0.0067$ and $p = 0.0031$), clearly indicating a role of MMP7 in cancer progression.

Although variations based on ER status and lack of statistical significance in differences between NATs and tumors of AA samples might be due to small sample size and should be interpreted with caution, Table 4 still suggests that genetic ancestry can play a critical role in defining overexpression in tumor samples. Additionally, we did not have enough samples to confirm the low MMP7 expression in tumors of Hispanic population; like other research using biobanks, this study was limited by less common minority participation in tissue donation.[35] Additional studies with larger number of samples, especially from minority women, are needed both to establish whether overexpression does vary by genetic ancestry and to define overexpression using genetic ancestry considerate values. The influence of ancestral lineage might also apply to biomarkers for other diseases. We have shown recently that genetic ancestry influences the expression levels of PD-1 and PD-L1 in tumor microenvironment and tumors, respectively,[15] which are used clinically for selecting patients for immunotherapy. Even hemoglobin A1c, a common clinical biomarker, shows racial differences in its relationship to circulating glucose concentrations.[36]

This study did have additional limitations worth noting. Developing MMP7 as a biomarker might ultimately prove difficult, given the number of significant covariates in a small sample size. In particular, there was not enough power to investigate the relationship between MMP7 and tumor stage. Another limitation of this study arose from incomplete information on some clinical data, such as tumor grade and breast cancer risk

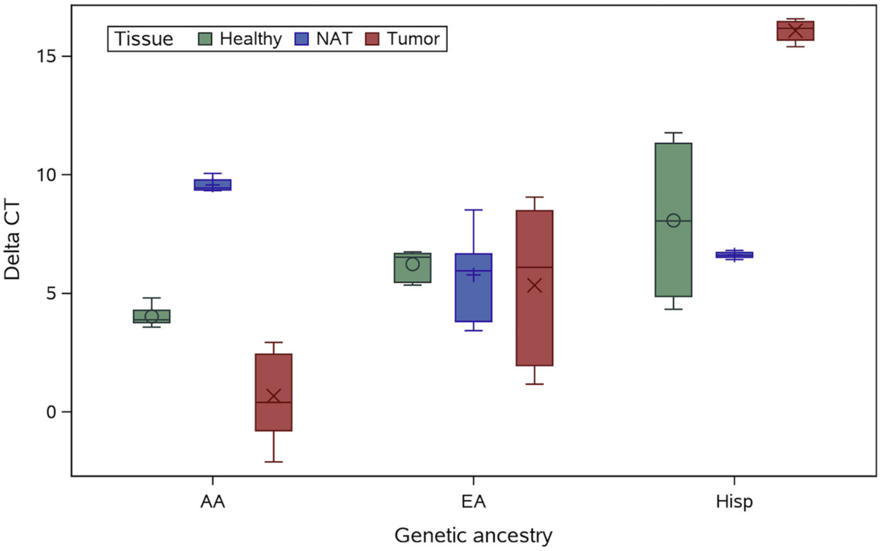


Figure 2. Quantitative reverse transcription polymerase chain reaction results for matrix metalloproteinase 7 (MMP7) expression in 9 tumor, 6 normal tissues adjacent to tumor (NATs), and 6 healthy breast tissue samples. ΔC_T values calculated using β-actin as the reference gene. Trends in messenger RNA expression levels are reverse to trends in ΔC_T (increase in ΔC_T correlates to decreased expression). AA, African American; EA, European American; Hisp, Hispanic.

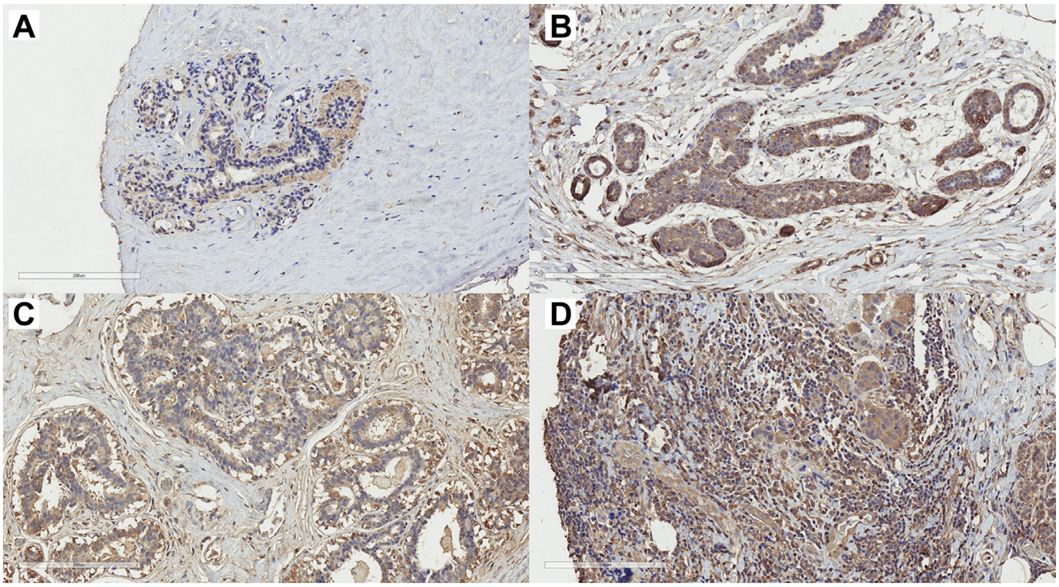


Figure 3. Immunohistochemistry staining with metalloproteinase 7 (MMP7). (A) Normal healthy breast sample with H-score of 19, (B) healthy breast sample with benign ductal hyperplasia with H-score of 67, (C) Normal adjacent to tumor sample with H-score of 105, and (D) tumor sample with H-score of 114.

factors. Future resources would likely need to focus on one subset of more homogenous patient samples at a time to tease apart the role of different clinical factors on MMP7 expression.

Cultured cells were used for initial RNA-sequencing screening and initial qRT-PCR validation, although clinically relevant biomarkers will not ultimately require cell culture. Cultured cells are an imperfect model that unfortunately can introduce discordance with measurements done on uncultured samples. However, given that healthy breast tissues are precious resources and, in our experience, cellularity of healthy tissue varies from 10% to 80%, screening for epithelial cell-enriched biomarkers with RNA sequencing on cultured epithelial cells was more practical. In addition, selection of luminal progenitor cells to reduce variability is achieved better in cell

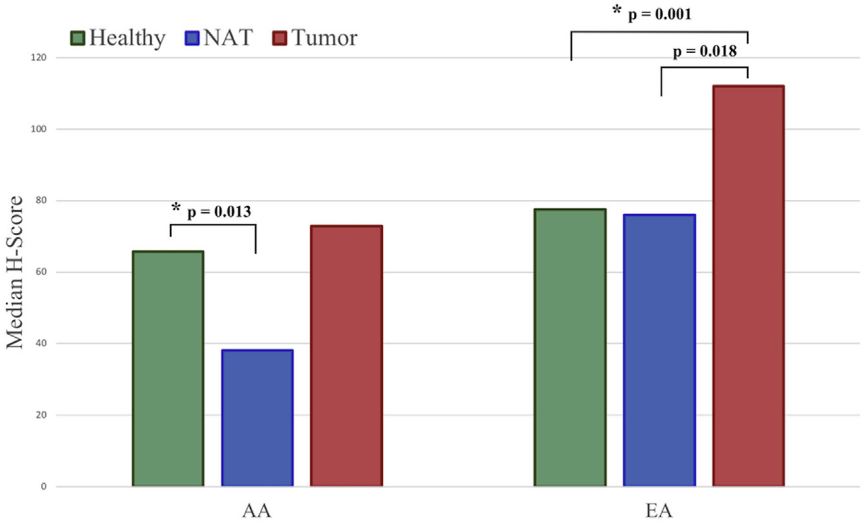


Figure 4. Immunohistochemistry staining for *MMP7* (median H-score) stratified by tissue type and genetic ancestry. Significant p values for H-score are shown. *p value is also < 0.05 for metalloproteinase 7 (MMP7) staining positivity. The median H-scores for healthy, normal tissues adjacent to tumor (NAT), and tumor samples were 66, 38, and 73 for African American (AA) samples and 78, 76, and 112 for European American (EA) samples.

Table 3. Matrix Metalloproteinase 7 Staining (H-score) Between Tumor and Normal Adjacent to Tumor Core Biopsy

	Tumor		Normal tissue adjacent to tumor		
Genetic ancestry and estrogen receptor status	**n**	**Median**	**n**	**Median**	**p Value**
African American					
Positive	19	60	6	37	0.6332
Negative	9	73	5	41	0.0234*
Total	31	73	11	38	0.0517
European American					
Positive	23	115	8	81	0.0445*
Negative	9	112	3	76	0.0645
Total	34	112	11	76	0.0181*
Combined	65	93	22	63	0.0067*

All p values calculated by 2-sided Wilcoxon test.
*Statistically significant.

culture. IHC analysis was used in part to overcome the limitation of cultured cells. In addition, concrete conclusions from qRT-PCR data must be cautioned, despite statistical significance due to its small sample size, and stratification was more appropriate with the larger sample size in the tissue microarrays used for IHC. Differences, however, between the protein and RNA studies suggest MMP7 expression is subject to regulation beyond mRNA expression. In this respect, mRNA and protein level correlation was observed with only one-third of RNA species and corresponding proteins examined in 23 cell lines.[37] Although the trend in tumor overexpression of MMP7 was reassuringly confirmed by IHC, the discordance in NAT MMP7 expression between mRNA assays and protein assays corroborates the conclusion that a biomarker identified at mRNA level, especially in cultured cells, needs to be verified at protein level, such as with IHC before additional development.

In consideration of the prolific body of evidence being built in the pursuit of novel biomarkers, our results can help to improve such searches. The breadth of avenues being pursued is encouraging, including recent studies on circulating cell-free nuclear DNA and mitochondrial DNA,[38] circular RNA,[39] circulating-tumor DNA,[40] and immunotherapy[12,13]; yet they have not incorporated genetic ancestry and often do not use true healthy normal as controls. Creating truly personalized cancer therapy will require more inclusive personalized data, such as genetic ancestry.

With respect to tumor biology and regulation, the mechanisms of MMP7 upregulation can perhaps be gleaned from earlier literature on colon and breast cancer. In colon cancer, Farnesoid X receptor, an intestinal tumor suppressor of unknown mechanism, represses MMP7 expression.[33] This same transcription factor protects gastric epithelial cells from inflammatory damage in mouse and human models.[41] Also, WNT5A increases

Table 4. Matrix Metalloproteinase 7 Staining (H-Score) Between Samples of African-American and European-American Genetic Ancestry

	African American		European American		
Tissue type and ER status	**n**	**Median**	**n**	**Median**	**p Value**
Tumor					
Total	31	73	34	112	0.0019*
ER+	19	60	23	115	0.0002*
ER−	9	73	9	112	0.1577
Normal tissue adjacent to tumor					
Total	11	38	11	76	0.0058*
ER+	6	37	8	81	0.0332*
ER−	5	41	3	76	NA†
Healthy	18	66	15	78	0.2701

All p values calculated by 2-sided Wilcoxon test.
*Statistically significant
†Too few patients to make comparison.
ER, estrogen receptor; NA, not applicable.

MMP7 via nuclear factor-κB signaling activation and contributes to the metastasis of FOXC1 overexpressing TNBC cells.[42] While Farnesoid X receptor, WNT5A, and FOXC1 expression was not significantly altered in epithelial cells of NATs or tumors compared with epithelial cells of healthy breast samples, Ingenuity Pathway Analysis linked 26 genes upregulated in NATs and tumors to nuclear factor-κB signaling network, which might have caused MMP7 upregulation.

CONCLUSIONS

Genomic differences across healthy, NAT, and tumor tissues in women with breast cancer demonstrate the presence of a para-inflammatory network in NATs. The combination of RNA sequencing, Ingenuity Pathway Analysis, qRT-PCR, and protein level analysis demonstrates that MMP7 expression is greater in tumors and is likely involved in the para-inflammatory network associated with breast cancer. Its expression in tumor-adjacent normal tissues, however, is significantly influenced by genetic ancestry and inter-individual differences and its investigation yields additional noteworthy lessons. This work suggests that use of healthy breast tissues instead of NATs as "normal" controls, combination of protein-based and transcriptome-based assays, and the incorporation of genetic ancestry in addition to traditional tumor subtyping, are all critical considerations that future investigators should use in developing meaningful biomarkers.

Author Contributions

Study conception and design: Nakshatri
Acquisition of data: Dirks, Anjanappa
Analysis and interpretation of data: Dirks, Burney, Sandusky, Hao, Liu, Nakshatri
Drafting of manuscript: Dirks
Critical revision: Schmidt, Nakshatri

Acknowledgment: The authors thank the members of the Indiana University Simon Cancer Center tissue procurement facility, flow cytometry core, Indiana University School of Medicine immunohistochemistry and Medical Genomics cores, and Komen Tissue Bank for their service and the countless number of women who donated their breast tissue for research purpose, as well as the volunteers who facilitated tissue collection.

REFERENCES

1. Gong M, Yan C, Jiang Y, et al. Genome-wide bioinformatics analysis reveals CTCFL is upregulated in high-grade epithelial ovarian cancer. Oncol Lett 2019;18:4030–4039.
2. Wu X, Han Y, Liu F, Ruan L. Downregulations of miR-449a and miR-145-5p act as prognostic biomarkers for endometrial cancer. J Comput Biol 2019 Sep 12 [Epub ahead of print].
3. Pattyn J, Van Keer S, Teblick L, et al. HPV DNA detection in urine samples of women: "An efficacious and accurate alternative to cervical samples?" Expert Rev Anti Infect Ther 2019 Sep 13 [Epub ahead of print].
4. Cozar JM, Robles-Fernandez I, Rodriguez-Martinez A, et al. The role of miRNAs as biomarkers in prostate cancer. Mutat Res 2019;781:165–174.
5. Khanmohammadi A, Aghaie A, Vahedi E, et al. Electrochemical biosensors for the detection of lung cancer biomarkers: a review. Talanta 2020;206:120251.
6. Le Y, Kan A, Li QJ, et al. NAP1L1 is a prognostic biomarker and contribute to doxorubicin chemotherapy resistance in human hepatocellular carcinoma. Cancer Cell Int 2019;19:228.
7. Thakur R, Laye JP, Lauss M, et al. Transcriptomic analysis reveals prognostic molecular signatures of stage I melanoma. Clin Cancer Res 2019 Sep 12 [Epub ahead of print].
8. Zheng K, Yang Q, Xie L, et al. Overexpression of MAGT1 is associated with aggressiveness and poor prognosis of colorectal cancer. Oncol Lett 2019;18:3857–3862.
9. Gerratana L, Davis AA, Shah AN, et al. Emerging role of genomics and cell-free DNA in breast cancer. Curr Treat Options Oncol 2019;20[8]:68.
10. Fares J, Kanojia D, Rashidi A, et al. Diagnostic clinical trials in breast cancer brain metastases: barriers and innovations. Clin Breast Cancer 2019;19:383–391.
11. Toss A, Venturelli M, Peterle C, et al. Molecular biomarkers for prediction of targeted therapy response in metastatic breast cancer: trick or treat? Int J Mol Sci 2017;18[1].
12. Adams S, Mittendorf EA. Lack of robust prognostic biomarkers for immunotherapy in breast cancer-adverse events-in reply. JAMA Oncol 2019 Sep 12 [Epub ahead of print].
13. Arora S, Velichinskii R, Lesh RW, et al. Existing and emerging biomarkers for immune checkpoint immunotherapy in solid tumors. Adv Ther 2019;36:2638–2678.
14. Beltran-Garcia J, Osca-Verdegal R, Mena-Molla S, Garcia-Gimenez JL. Epigenetic IVD tests for personalized precision medicine in cancer. Front Genet 2019;10:621.
15. Nakshatri H, Kumar B, Burney HN, et al. Genetic ancestry-dependent differences in breast cancer-induced field defects in the tumor-adjacent normal breast. Clin Cancer Res 2019;25:2848–2859.
16. Teschendorff AE, Gao Y, Jones A, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun 2016;7:10478.
17. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–674.
18. Aran D, Lasry A, Zinger A, et al. Widespread parainflammation in human cancer. Genome Biol 2016;17:145.
19. Todoric J, Karin M. The fire within: cell-autonomous mechanisms in inflammation-driven cancer. Cancer Cell 2019;35:714–720.
20. Nakshatri H, Anjanappa M, Bhat-Nakshatri P. Ethnicity-dependent and -independent heterogeneity in healthy normal breast hierarchy impacts tumor characterization. Sci Rep 2015;5:13526.
21. Raouf A, Zhao Y, To K, et al. Transcriptome analysis of the normal human mammary cell commitment and differentiation process. Cell Stem Cell 2008;3:109–118.
22. Lim E, Vaillant F, Wu D, et al. Aberrant luminal progenitors as the candidate target population for basal tumor

development in BRCA1 mutation carriers. Nat Med 2009;15: 907–913.

23. Proia TA, Keller PJ, Gupta PB, et al. Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. Cell Stem Cell 2011;8:149–163.

24. Liu X, Ory V, Chapman S, et al. ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells. Am J Pathol 2012;180:599–607.

25. Prasad M, Kumar B, Bhat-Nakshatri P, et al. Dual TGFbeta/BMP pathway inhibition enables expansion and characterization of multiple epithelial cell types of the normal and cancerous breast. Mol Cancer Res 2019;17:1556–1570.

26. Visvader JE, Stingl J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes Dev 2014;28:1143–1158.

27. Kumar B, Prasad M, Bhat-Nakshatri P, et al. Normal breast-derived epithelial cells with luminal and intrinsic subtype-enriched gene expression document interindividual differences in their differentiation cascade. Cancer Res 2018;78: 5107–5123.

28. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 2001;25:402–408.

29. Goswami CP, Nakshatri H. PROGgeneV2: enhancements on the existing database. BMC Cancer 2014;14:970.

30. Vreemann S, Gubern-Merida A, Lardenoije S, et al. The frequency of missed breast cancers in women participating in a high-risk MRI screening program. Breast Cancer Res Treat 2018;169:323–331.

31. Lappalainen T, Sammeth M, Friedlander MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013;501[7468]:506–511.

32. Alaseem A, Alhazzani K, Dondapati P, et al. Matrix metalloproteinases: a challenging paradigm of cancer management. Semin Cancer Biol 2019;56:100–115.

33. Peng Z, Chen J, Drachenberg CB, et al. Farnesoid X receptor represses matrix metalloproteinase 7 expression, revealing this regulatory axis as a promising therapeutic target in colon cancer. J Biol Chem 2019;294:8529–8542.

34. Sizemore ST, Sizemore GM, Booth CN, et al. Hypomethylation of the MMP7 promoter and increased expression of MMP7 distinguishes the basal-like breast cancer subtype from other triple-negative tumors. Breast Cancer Res Treat 2014;146:25–40.

35. Kim P, Milliken EL. Minority participation in biobanks: an essential key to progress. Methods Mol Biol 2019;1897: 43–50.

36. Bergenstal RM, Gal RL, Connor CG, et al. Racial differences in the relationship of glucose concentrations and hemoglobin A1c levels. Ann Intern Med 2017;167:95–102.

37. Gry M, Rimini R, Stromberg S, et al. Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genomics 2009;10:365.

38. Pasha HA, Rezk NA, Riad MA. Circulating cell free nuclear DNA, mitochondrial DNA and global DNA methylation: potential noninvasive biomarkers for breast cancer diagnosis. Cancer Invest 2019;37:432–439.

39. Li Z, Chen Z, Hu G, Jiang Y. Roles of circular RNA in breast cancer: present and future. Am J Transl Res 2019;11:3945–3954.

40. Hironaka-Mitsuhashi A, Sanchez Calle A, Ochiya T, et al. Towards circulating-tumor DNA-based precision medicine. J Clin Med 2019;8[9].

41. Lian F, Xing X, Yuan G, et al. Farnesoid X receptor protects human and murine gastric epithelial cells against inflammation-induced damage. Biochem J 2011;438: 315–323.

42. Han B, Zhou B, Qu Y, et al. FOXC1-induced non-canonical WNT5A-MMP7 signaling regulates invasiveness in triple-negative breast cancer. Oncogene 2018;37:1399–1408.

Discussion

DR KELLY McMASTERS (Louisville, KY): Primary screen was performed by conducting RNA sequencing of primary cells grown in culture and sorted by flow cytometry from 12 breast tissue samples, both from estrogen receptor (ER)-positive and -negative tumors in patients with varied genetic ancestry. This formed the basis for all subsequent analyses.

The more variables introduced at the front end of the screening process, the more variability you will get at the end. If you had this to do over again, would you simplify this analysis in matched patient samples from a more homogeneous patient cohort to reduce the potential variability? You identified 26 unique genes with differential expression, yet you presented polymerase chain reaction (PCR) and protein data only for *MMP7*. The PCR validation was performed in cultured cells, but not sorted by flow cytometry. However, any useful biomarker will not require tissue culture of patient samples, so what about the other 25 genes? Were they confirmed by PCR, and did you try to study noncultured tissue samples by PCR?

In the evaluation of *MMP7* protein expression by tissue microarrays, you had to torture the data pretty hard from a pretty small sample size in order to find statistically significant differences in some subgroups based on ancestry and hormone receptor status, and there is no true validation cohort for this study. You even conclude in the manuscript that *MMP7* is likely not a meaningful biomarker for breast cancer. So, based on your experience with this study, and a lot of hard work and data, how do you plan to use these data and modify your approach to biomarker discovery in the future? In the end, I think the authors have demonstrated for all of us the immense challenges of identifying and validating unique clinically relevant and useful cancer biomarkers.

DR C MAX SCHMIDT (Indianapolis, IN): In answer to your first question about simplifying this through a more homogeneous patient cohort, we did select for only luminal progenitor cells for RNA sequencing, thinking that this would help reduce variability initially, and we did do a screen with normal adjacent tissue and tumor to avoid bias. But if we did it again, we would select all European, American, and ER-positive patients to further reduce variability. We suspect, though, that when genetic ancestry and hormonal status are added later, the same incongruity would occur.

To address your second question, regarding starting with noncultured tissue samples, we wanted to again focus on a less variable selection of luminal progenitor cells, sorting to achieve better cell