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

**Use Healthy Controls**

**Consider Genetic Ancestry**

**Use Transcriptomic & Protein Assays**

Key Considerations for Developing Breast Cancer Biomarkers: Lessons from MMP7

Dirks et al. J Am Coll Surg, July 2020
Breast Heterogeneity: Obstacles to Developing Universal Biomarkers of Breast Cancer Initiation and Progression

Rebecca C Dirks, MD\textsuperscript{a}, Heather N Burney, MS\textsuperscript{b}, Manjushree Anjanappa, MS\textsuperscript{a}, George E Sandusky, DVM, PhD\textsuperscript{c}, Yangyang Hao, PhD\textsuperscript{d}, Yunlong Liu, PhD\textsuperscript{d}, Max C Schmidt, MD, PhD\textsuperscript{a}, Harikrishna Nakshatri, BVSc, PhD\textsuperscript{a,e,f}

\textsuperscript{a}Department of Surgery, Indiana University School of Medicine, Indianapolis, IN 46202, USA
\textsuperscript{b}Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
\textsuperscript{c}Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA
\textsuperscript{d}Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
\textsuperscript{e}Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA
\textsuperscript{f}Roudebush VA Medical Center, Indianapolis, IN 46202, USA

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 and 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.

Corresponding Author:
Harikrishna Nakshatri, B.V.Sc., PhD.
C218C, 980 West Walnut St.
Indianapolis, IN 46202, USA
317 278 2238 (phone)
317 274 0396 (fax)
hnakshat@iupui.edu

Brief Title: Biomarkers of Breast Cancer Progression
**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 further using normal tissue adjacent to tumor (NATs) with 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 PCR (qRT-PCR) 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 to healthy controls at mRNA level and formed a para-inflammatory network. Matrix metalloproteinase-7 (MMP7) had the greatest expression in tumor cells, with upregulation confirmed by qRT-PCR. Tumor-enriched but not NAT-enriched expression of MMP7 compared to 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.

**Conclusion:** 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.
Keywords: breast cancer, biomarker, genetic ancestry, para-inflammation, MMP7
Abbreviations

AA - African American
CD49f - CD49 Antigen-Like Family Member F
EA - European American
EpCAM - Epithelial cell adhesion molecule
ER - estrogen receptor
IHC - Immunohistochemistry
MMP7 - Matrix metalloproteinase 7, Matrilysin
mRNA – messenger RNA
NAT - normal adjacent (to tumor) tissue
qRT-PCR - quantitative reverse transcription polymerase chain reaction
Introduction

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). While 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 may hold a key to early cancer recognition. Non-malignant, 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). While not malignant, reprogrammed epithelial cells can contribute to tumor initiation/progression (19) and para-inflammatory changes within NATs could serve as
biomarkers. Prior 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 utilize 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, NATs and tumors 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), over 2000 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 MMP7, a member of the matrix metalloproteinase family of zinc dependent endopeptidases, for further exploration at mRNA and protein levels.

**Methods**

Primary cell lines and culture

Breast epithelial cells from de-identified healthy tissue were obtained from the Komen Tissue Bank at Indiana University, containing core biopsies donated from healthy women. Samples with diverse genetic ancestry were sought for inclusion. De-identified tumor tissues and NATs were obtained from surgical cases at Indiana University after prior 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 qRT-PCR, we employed an improved method developed in the lab, which does not require irradiated mouse embryonic fibroblast feeder layer. Briefly, this method utilized culture dishes precoated with laminin-5-rich-conditioned media from 804G cell line and a growth media supplemented with inhibitors of Rho Kinase (ROCK), transforming growth factor beta (TGFbeta), and bone morphogenetic protein (BMP) signaling (25).

**RNA sequencing and Ingenuity Pathway Analysis**

RNA sequencing was performed on primary cells from healthy, NAT, and tumor cells. All primary cells were sorted by flow cytometry prior to sequencing to enrich for luminal progenitor cells. Combination CD49f positive and EpCAM positive (CD49f+/EpCAM+) 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/EpCAM staining patterns of three samples from healthy breast, three paired NATs and tumors and DCIS and LCIS 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 cDNA library was created for sequencing. RNA sequencing was performed as previously described (27). Differential expression analysis was accomplished using edgeR package and both p values and false discovery rate were computed. Ingenuity
Pathway Analysis (Qiagen) was employed 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.

**qRT-PCR validation**

Primary cells were cultured without prior selection by flow cytometry, and RNA from exponentially growing cells was isolated with RNeasy Kit (74106, Qiagen). cDNA 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 quantitative real-time PCR (qRT-PCR). For matched pairs of NAT and tumor cells, analysis of resultant qRT-PCR data was performed using the double delta $C_T$ analysis described by Livak and Schmittgen (28). Beta-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 non-paired samples, $\Delta C_T$ 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 and European American 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 HIPAA protocol. A
CLIA-certified pathology core was used for immunohistochemistry and three blinded pathologists employed 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 were removed from the database and labeled on provided excel master database. Additionally, any cores with extensive physical damage (tears, folds) were excluded from analysis.

**Statistical Analysis**

All experiments were conducted in at least three biological 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 Statistical Analysis Software (SAS) University Edition version 2018 (Cary, North Carolina). 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 alpha value of 0.05 was set for statistical significance (p < 0.05, false discovery rate < 0.05). MMP-7 positivity and H-scores were analyzed by non-parametric Wilcoxon rank-sum tests. All statistically significant differences are described in the text while acknowledging that small sample size may have affected these analyses in some instances.

**Results**
Inflammation-associated genes are upregulated in NATs and tumors compared to 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 European American (EA), 1 Hispanic (Hisp), and 2 African American (AA) samples. EA samples included two matched NAT-tumor pairs. Hormone receptor status was known for tumor samples. Except for one NAT sample, all RNA-sequencing samples were from estrogen receptor alpha-positive (ER+) and progesterone receptor positive (PR+) breast cancer. None were HER2 positive (human epidermal growth factor receptor 2). Stage, nodal status, and grade were available for all tumor samples and all but one 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 may 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, 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 to 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, two networks were generated with 99% confidence that a similar network could not be created by random chance. These networks had three 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-seq data of MMP7 by qRT-PCR**

RNA sequencing results for MMP7 were validated in non-sorted 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-seq data, suggesting greater relevance as a breast cancer marker.

MMP7 qRT-qPCR was performed on the 2 paired tumor-NAT samples from RNA seq (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 (Figure 1B), similar to RNA-seq 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 one 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 Hisp), 6 NAT samples (4 EA, 1 AA, 1 Hisp), and 6 healthy samples (2 each EA, AA, and Hisp). The healthy samples included 5 Komen Tissue Bank samples and one contralateral prophylactic breast tissue sample.

When all samples were tested for MMP7 overexpression, there was a statistically significant difference in $\Delta C_T$ values across healthy, NAT, and tumor samples ($p = 0.0035$). In multivariable analysis, tissue type, genetic ancestry, and the interaction between these two variables (tissue*genetic ancestry interaction) all significantly influenced MMP7 level ($p = 0.014$, $p < 0.001$, $p < 0.001$, respectively). This difference across both variables can be seen in Figure 2 where $\Delta C_T$ is inversely proportional to mRNA expression. While tumors of AA women, and to some extent EA women, expressed higher levels of MMP7 compared to NATs or healthy, tumors of Hispanic women contained lower levels of MMP7. These results were approached with caution given the limited sample size, especially from Hispanic and African American 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 HER2 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 4 samples having greater ΔC_T values, and thus smaller MMP7 expression than stage 1 or 2 (p < 0.001). There were no stage 3 tumor samples cultured for qRT-PCR.

**Discordance between MMP7 transcripts and protein in normal breasts, NATs and tumors:**

A tissue microarray with 67 tumors, 23 NATs, and 33 healthy breasts undamaged cores was used for 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 to healthy breast (Figure 4). Surprisingly, NATs of AA women had lower levels of MMP7 compared to healthy counterparts (p = 0.0126), directly opposite to the trend in mRNA in RNA-seq. However, qRT-PCR and IHC data are compatible (Figure 2 and Figure 4) with lower expression in NATs compared to healthy tissues of AA women. Additionally, MMP7 H-score was still statistically increased in all tumor samples compared to NATs and to 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 to 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 0-2 years before cancer detection had MRI detectable cancers that were missed, whereas 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 prior 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-seq 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 three 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 to healthy breasts for further analyses.
While networks created from Ingenuity Pathway Analysis of the 26 genes involved in 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 to a more appropriate control (healthy samples from the Komen Tissue Bank). Since 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 matrix metalloproteinase family of zinc dependent endopeptidases (MMPs), MMP7 and MMP12, were notable members of Ingenuity Pathway Analysis inflammatory networks of NATs. This corroborates prior literature concerning matrix metalloproteinases, inflammation, and cancer. As summarized in a 2017 review on MMPs by Alaseem et al, matrix metalloproteinases have an established connection with malignancy, including proliferation, invasion, angiogenesis and metastasis, and modulate inflammatory events to worsen pathological conditions (32). While 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 matrix metalloproteinase 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 further investigated.
Genetic ancestry can influence the field effects found in NATs. For example, we have previously reported genetic ancestry may 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 3-4 and Figures 1-2,4). 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 to NATs and to healthy samples (p = 0.0067, p = 0.0031), clearly indicating a role of MMP7 in cancer progression.

While variations based on ER status and lack of statistical significance in differences between NATs and tumors of AA samples may 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 may also apply to biomarkers for other diseases. We have shown recently 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 A_1c, 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 may 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 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-seq screening and initial qRT-PCR validation although clinically relevant biomarkers ultimately won’t require cell culture. Cultured cells are an imperfect model that unfortunately may introduce discordance with measurements done on uncultured samples. However, given healthy breast tissues are precious resources, and, in our experience, cellularity of healthy tissue varies from 10- 80%, screening for epithelial cell-enriched biomarkers with RNA-seq on cultured epithelial cells was more practical. Further, selection of luminal progenitor cells to reduce variability is achieved better in cell 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, whereas 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 a third of RNA species and corresponding proteins examined in 23 cell lines (39). 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 further 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 further 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 nDNA and mitochondrial DNA (mtDNA)(40), circular RNA(41), circulating-tumor DNA (ctDNAs) (42), 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 prior literature on colon and breast cancer. In colon cancer, Farnesoid X receptor (FXR), 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 (37). Also, WNT5A increases MMP7 via NF-κB signaling activation and contributes to the metastasis of FOXC1 overexpressing TNBC cells (38). While FXR, WNT5A, and FOXC1 expression was not significantly altered in epithelial cells of NATs or tumors compared to epithelial cells of healthy breast samples, Ingenuity Pathway Analysis linked 26 genes upregulated in NATs and tumors to NF-κB signaling network, which may have caused MMP7 upregulation.

**Conclusion**

Genomic differences across healthy, NATs, 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-inflamatory 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 future investigators should use in developing meaningful biomarkers.

Acknowledgements: We thank the members of IU Simon Cancer Center tissue procurement facility, flow cytometry core, IU 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 volunteers who facilitated tissue collection.
References


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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Average CPM</th>
<th>logFC</th>
<th>Paired tumor vs NAT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tumor (n = 5)</td>
<td>NAT (n = 3)</td>
<td>Healthy (n = 4)</td>
</tr>
<tr>
<td>MMP7</td>
<td>40.949</td>
<td>4.164</td>
<td>0.801</td>
</tr>
<tr>
<td>PTPRJ</td>
<td>18.787</td>
<td>4.479</td>
<td>1.161</td>
</tr>
<tr>
<td>STAB1</td>
<td>11.575</td>
<td>4.073</td>
<td>0.853</td>
</tr>
<tr>
<td>REN</td>
<td>10.956</td>
<td>3.072</td>
<td>1.173</td>
</tr>
<tr>
<td>GATM</td>
<td>8.064</td>
<td>1.834</td>
<td>0.683</td>
</tr>
<tr>
<td>CYBB</td>
<td>7.065</td>
<td>1.612</td>
<td>0.510</td>
</tr>
<tr>
<td>VTRNA1-2</td>
<td>6.767</td>
<td>3.361</td>
<td>1.302</td>
</tr>
<tr>
<td>EMR1</td>
<td>6.273</td>
<td>1.347</td>
<td>0.180</td>
</tr>
<tr>
<td>ITGAM</td>
<td>6.056</td>
<td>1.670</td>
<td>0.178</td>
</tr>
<tr>
<td>PTPRC</td>
<td>3.110</td>
<td>0.874</td>
<td>0.250</td>
</tr>
<tr>
<td>HCK</td>
<td>3.057</td>
<td>0.796</td>
<td>0.177</td>
</tr>
<tr>
<td>CP</td>
<td>2.415</td>
<td>0.460</td>
<td>0.013</td>
</tr>
<tr>
<td>TNFRSF11B</td>
<td>2.071</td>
<td>0.452</td>
<td>0.146</td>
</tr>
<tr>
<td>TLR8</td>
<td>1.996</td>
<td>0.587</td>
<td>0.161</td>
</tr>
<tr>
<td>SLC11A1</td>
<td>1.779</td>
<td>0.607</td>
<td>0.159</td>
</tr>
<tr>
<td>MARCO</td>
<td>1.740</td>
<td>0.413</td>
<td>0.008</td>
</tr>
<tr>
<td>TBXAS1</td>
<td>1.711</td>
<td>0.552</td>
<td>0.147</td>
</tr>
<tr>
<td>C4orf26</td>
<td>1.688</td>
<td>0.207</td>
<td>0.029</td>
</tr>
<tr>
<td>IRG1</td>
<td>1.340</td>
<td>0.085</td>
<td>0.000</td>
</tr>
<tr>
<td>NEK5</td>
<td>1.281</td>
<td>0.130</td>
<td>0.012</td>
</tr>
<tr>
<td>MMP12</td>
<td>0.936</td>
<td>0.169</td>
<td>0.037</td>
</tr>
<tr>
<td>CFTR</td>
<td>0.872</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td>ITGAD</td>
<td>0.784</td>
<td>0.148</td>
<td>0.028</td>
</tr>
<tr>
<td>CD207</td>
<td>0.630</td>
<td>0.079</td>
<td>0.000</td>
</tr>
<tr>
<td>CCL2</td>
<td>0.525</td>
<td>0.125</td>
<td>0.020</td>
</tr>
<tr>
<td>BTK</td>
<td>0.501</td>
<td>0.116</td>
<td>0.012</td>
</tr>
</tbody>
</table>

All fold changes p value < 0.001 and false discovery rate < 0.05
* false discovery rate < 0.001
CPM, counts per million; FC, fold change; NAT, normal adjacent to tumor
Table 2. Ingenuity Pathway Analysis Networks and Associated Top Diseases and Functions

<table>
<thead>
<tr>
<th>Network, shared top disease and function</th>
<th>Included gene from RNA-sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADGRE1, BTK, CCL2, CFTR, HCK, ITGAM, MMP7, MMP12, PTPRC, PTPRJ, SLC11A1, TLR8, TNFRSF11B</td>
</tr>
<tr>
<td>Cellular movement, immune cell trafficking, inflammatory response, cellular function and maintenance</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CP, CYBB, ITGAD, ODAPH, REN, STAB1, VTRNA1-2</td>
</tr>
<tr>
<td>Hematological system development and function, inflammatory response, tissue morphology</td>
<td></td>
</tr>
<tr>
<td>Connective tissue development and function, lipid metabolism, small molecule biochemistry</td>
<td>CP, CYBB, ITGAD, ODAPH, REN, STAB1, VTRNA1-2</td>
</tr>
</tbody>
</table>
Table 3. *MMP7* Staining (H-score) between Tumor and Normal Adjacent to Tumor Core Biopsy

<table>
<thead>
<tr>
<th>Genetic ancestry and ER status</th>
<th>Tumor</th>
<th>NAT</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Median</td>
<td>n</td>
</tr>
<tr>
<td><strong>African American</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>19</td>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>Negative</td>
<td>9</td>
<td>73</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>73</td>
<td>11</td>
</tr>
<tr>
<td><strong>European American</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>23</td>
<td>115</td>
<td>8</td>
</tr>
<tr>
<td>Negative</td>
<td>9</td>
<td>112</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>112</td>
<td>11</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td>65</td>
<td>93</td>
<td>22</td>
</tr>
</tbody>
</table>

All p values calculated by 2-sided Wilcoxon test

*Statistically significant

ER, estrogen receptor; NAT, normal adjacent to tumor
Table 4. MMP7 staining (H-score) between samples of African American and European American Genetic Ancestry

<table>
<thead>
<tr>
<th>Tissue type and ER status</th>
<th>African American</th>
<th>European American</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Median</td>
<td>n</td>
</tr>
<tr>
<td>Tumor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>73</td>
<td>34</td>
</tr>
<tr>
<td>ER+</td>
<td>19</td>
<td>60</td>
<td>23</td>
</tr>
<tr>
<td>ER-</td>
<td>9</td>
<td>73</td>
<td>9</td>
</tr>
<tr>
<td>NAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>38</td>
<td>11</td>
</tr>
<tr>
<td>ER+</td>
<td>6</td>
<td>37</td>
<td>8</td>
</tr>
<tr>
<td>ER-</td>
<td>5</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>Healthy</td>
<td>18</td>
<td>66</td>
<td>15</td>
</tr>
</tbody>
</table>

All p values calculated by two-sided Wilcoxon test
*Statistically significant
†Too few patients to make comparison
ER, estrogen receptor; N/A, not applicable; NAT, normal adjacent to tumor
Figure Legends

**Fig 1.** Characterization of epithelial cells derived from healthy breast, normal adjacent to tumor (NATs) and tumors. A) CD49f/EpCAM 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 *MMP7* expression in 2 paired NAT-tumor samples. Relative fold change was calculated using double ΔC_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 ≤ 0.001, 2-way ANOVA) and a median fold change of 2.2 for woman B (*p > 0.05, 2-way ANOVA).

**Fig 2.** Quantitative reverse transcription polymerase chain reaction results for *MMP7* expression in 9 tumor, 6 normal adjacent to tumor (NAT), and 6 healthy breast tissue samples. ΔC_T values calculated using β-Actin as the reference gene. Trends in mRNA 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.

**Fig 3.** Immunohistochemistry staining with *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.

**Fig 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 *MMP7* staining positivity. The median H-scores for healthy, normal 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.
Precis

This study investigates para-inflammation-associated biomarkers for breast cancer initiation and progression. It illustrates the complexity in developing universal biomarkers, which is most likely due to genetic diversity in the human population.
Figure 1

A

NAT-1

EpCAM-PE

Luminal progenitors

Feeder layer fibroblasts

NAT-2

Tumor

Paired sample

NAT

Tumor

Paired sample

DCIS

LCIS

Same patient

B

CD49f-APC

CD49f-APC

Fold Change

NAT  Tumor

A  Sample  B  Figure 1