5-2022

Identifying Genetic Differences Among African American And Caucasian Triple Negative Breast Cancer Genotypes

Christopher Jordan Dixon

IDENTIFYING GENETIC DIFFERENCES AMONG AFRICAN AMERICAN AND CAUCASIAN TRIPLE NEGATIVE BREAST CANCER GENOTYPES

THESIS

Presented in Partial Fulfillment of the Requirements for

the Degree Master of Science in the Graduate School

of Texas Southern University

By

Christopher Jordan Dixon, B.S.

Texas Southern University

2022

Approved by

Dr. Audrey Player
Chairperson, Thesis Committee

Dr. Gregory Maddox
Dean, The Graduate School

Approved by

Dr. Audrey Player	03/25/2022
Chairperson, Thesis Committee	Date
Dr. Mario Hollomon	03/25/2022
Committee Member	Date
Dr. Tuan D Phan	03/25/2022
Committee Member	Date
Dr. Ayodotun Sodipe	03/25/2022
Committee Member	Date

DETERMINING THE GENETIC DIFFERENCES BETWEEN AFRICAN AMERICAN AND CAUCASIAN TRIPLE NEGATIVE BREAST CANCER GENOTYPES

By

Christopher Jordan Dixon, M.S.

Texas Southern University, 2022

Associate Professor Audrey N. Player, Ph.D., Advisor

Triple negative breast cancers (TNBC) are closely related to basal-like cancers and classified based on their molecular signatures and their progenitor cell type. TNBCs lack the presence of three common types of receptors known to fuel breast cancer growth: estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptors 2 (HER2neu). TNBC represent 10-20% of all molecular breast cancer subtypes. Even though genomic and transcriptome analyses show that many of the molecular signatures associated with TNBC are not related to ethnicity, clinicians and researchers find that African American (AA) TNBC women have higher mortality rates compared to Caucasian (CA) women. The high mortality rates are linked to socioeconomic factors like access to adequate healthcare, but researchers are exploring the possibility that genetic differences between AA and CA patients may also play a role in racial disparities. Microarray analyses have been instrumental in characterizing TNBC and many other types of breast cancer. Related to TNBC, microarray analyses (a) validate the negative receptor-

status of the cancers (b) identify and define the six sub-categories of TNBC validating the heterogeneity of the cancers as defined by Lehmann et al and (c) the microarray gene expression platform is proving to be useful towards determining genes differentially represented in AA and CA TNBC. Our approach is to use the microarray platform (and a cell line model) to further examine the differences between the transcriptomes of CA and AA women. For more accurate transcriptome comparisons, we've identified and compared *AA Basal-A TNBC* to *CA Basal-A TNBC*, and separately *AA Basal-B TNBC* compared to *CA Basal-B TNBC*. Bioinformatic analyses show that TCEAL8 and TCEAL9 genes, both located on X-chromosome are differentially expressed in AA compared to CA TNBC. The EFHD1 gene is identified as differentially expressed in *AA Basal-B* compared to *CA Basal-B TNBC*. These data serve as a preliminary study towards further characterizing molecular differences between the transcriptomes of AA compared to CA TNBC patient populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2Q37	Breast Cancer Susceptibility Allele
AA	African American
APOE	Apoliipoprotein E
BCRA1	Breast Cancer Type 1
BEX	Brain Expressed X-linked
BLAST	Basic Local Alignment Search Tool
CA	Caucasian
CRC	Colorectal Cancer
DNA	Deoxyribonucleic Acid
EFHD1	EF Hand Domain 1
ER	Estrogen Receptor
GDS3097	Inflammatory Breast Cancer Cell Line
HCC1143	B Lymphoblastoid Cell Line
HCC1187	B Lymphoblastoid Cell Line
HCC70	Epithelial Cell Line
HER 2	Human Epidermal Growth Factor 2
HER+ve	Human Epidermal Growth Factor 2 Positive
mAdb	NCI/CIT Microarray Database
MB468	Cell line isolated from 51-year-old female in 1977
MDA-MB157	Epithelial Cell Line
MDA-MB231	Epithelial Cell Line
MDA-MB 436	Human Breast Carcinoma Cell Lines

MRI	Magnetic Resonance Imaging
NIH	National Institute of Health
P value	Measure of the probability an observed difference could occur
PEG	Paternally Expressed Gene
PR	Progesterone Receptor
RNA	Ribonucleic Acid
T-test	Statistical Hypothesis Test
TCEAL	Transcription Elongation Factor
TNBC	Triple Negative Breast Cancer
WHO	World Health Organization

VITA

2008 …….……	Louis Stokes Alliance for Minority Participation Internship University of Texas at Dallas Dallas, Texas
2009 …….……	B.S. Biology/Chemistry Minor Wiley College, Marshall, Texas
2009 …….	Human Genome Sequencing Center-Genetics/Genomics University of Texas at Dallas Dallas, Texas
2018 …….	Graduate Texas Assistant Department of Biology Texas Southern University Houston, Texas
2022 ……	National Association of African American Studies Presenter
Major Field……	Biology

ACKNOWLEDGEMENTS

I, Christopher Jordan Dixon, would like to thank and give all honor and praise to my Lord and Savior Jesus Christ for guidance and wisdom in completing this project. I would like to express my sincere gratitude to my advisor and mentor, Dr. Audrey Player, for entrusting in me the task to take on a project of this magnitude. Dr. Audrey Player has always been encouraging, motivating and patient with me throughout this project, and I want to thank her for her brilliant input and guidance, which largely aided in getting this project completed. I also would like to give my sincere thanks to my thesis committee Dr. Ayodotun Sodipe, Dr. Mario Hollomon and Dr. Tuan Phan for their contributions of their time and input during the completion of this project.

Thank you to Texas Southern University for their acceptance of me into the master's program, and a special thanks to the College of Science, Engineering and Technology Biology department for their support during my master's matriculation. I would like to greatly thank Dr. Debra Murray, Professor of Molecular and Human Genetics at Baylor College of Medicine for encouraging me to apply to Texas Southern University and always being available to offer advice and immense knowledge while successfully navigating my way through the science field. Thanks to all my colleagues, especially my lab partners, Esther Nsende, Tyra Ivory and Sierra Cunningham for their dedicated contributions to Dr. Audrey Player's lab and for making my matriculation an enjoyable experience. I must also show my extreme gratitude to my father and mother, Gery and Dianne Dixon, and my brother and sister, Gery Dixon Jr. and Courtney Dixon, for their continued support as I pursue my goals in higher learning.

CHAPTER 1

INTRODUCTION

Cancer

Cancer is defined as a disease where cells undergo differentiation, leading to uncontrollable growth and in some cases metastases to distinct locations in the body (1). The human body is made up of trillions of cells, all of which can mutate and ultimately differentiate into cancerous cells (1). Mutations within a cell can activate oncogenes or deactivate tumor suppressor genes that affect differentiation, cell cycle signaling pathways, cell death mechanisms and yet undefined processes (2). The cancerous cell proliferates, continues to differentiate, and ultimately becomes invasive based on its genotype (3). The cancers become harmful when their growth interrupts the natural survival functions of the primary and metastatic organs (3).

Cancers develop into invasive tumors in stages (Figure 1). The first stage of development is hyperplasia, which is defined as an increase in the size of the tissue caused by an increase in cellular proliferation (3), which itself is not a cancer. The off spring of these cells then grow and differentiate, displaying a 'somewhat abnormal' phenotype and genotype; this stage is called dysplasia (3). Dysplastic cells that continue to mutate and stay within the tissue of origin are called in-situ tissues (3). In-situ tissues can remain in the site from which they originate, but if they continue to mutate and convert to cancers, shed into the blood and invade nearby tissues, they are described as malignant invasive

cancers (3). To determine the progression of cancers, pathologists will also describe a numeric 'staging' for the cancers as listed below:

Cancer Number Staging:

Stage 0- in situ, abnormal tissue localized to site of origin and has not spread

Stage 1-small cancer, that has not spread beyond its original boundaries

Stage 2-larger cancer, that has not spread beyond its original boundaries

Stage 3-larger cancer, that may have spread to the surrounding lymph nodes

Stage 4-the cancer has spread (i.e., metastasized to secondary organs

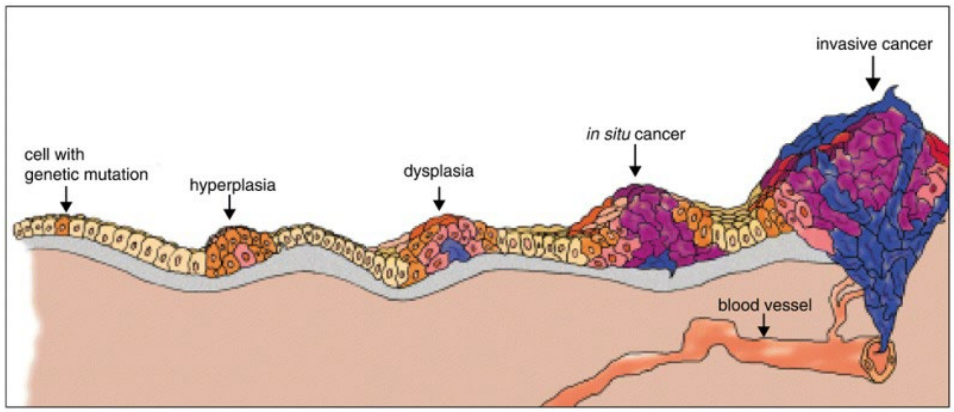


Figure 1: Tumor Development Stages
Tumor cells divide in stages which can ultimately invade other tissues or shed into the blood and travel throughout the body establishing tumors in other areas (8).

Common Types of Cancer

In 2020 alone, cancer caused nearly 10 million deaths making it the leading cause of death worldwide (4). The most common types of cancer are breast, lung, colon, prostate cancers and skin (4). **Breast cancer** is one of the most common types of cancer in women and leads all other types of cancers in cancer death rate among women (4). Over 200 types

of cancer have been identified but of these types of cancers breast cancer was the leading cancer diagnosed in the United States (5). The focus of my research is to better characterize Triple Negative Breast Cancer (TNBC) a type of breast cancer that has higher fatality rates in women of African descent. The different types of breast cancer will be discussed in the following sections. Lung cancer is a form of cancer that involves the formation of cancer cells in the lung tissue. It is the second most diagnosed cancer in the United States and causes the most deaths of all other types of cancers making up 25% of all cancer deaths (6). Colon cancer or Colorectal cancer (CRC) is the third most diagnosed cancer in the United States and is the second deadliest cancer type among men and women (7). CRC has steadily declined in individuals over 50, mainly due to cancer screening and therapeutic advancements (7). Prostate cancer is the most common type of cancer diagnosed in men, worldwide, and is the fourth highest cause of cancer death in men (8). Most prostate cancers tend to be less aggressive and grow much slower than other types of malignancies, and when limited to the prostate these cancers are generally considered localized and curable (8). Related to Skin Cancer, approximately 3.6 million new patients of Basal Cell Carcinoma, 1.8 million new patients of Squamous cell carcinoma, 197,000 new patients with Melanoma and 3000 new cases of Merkel Skin cancer are diagnosed world-wide each year (https://www.skincancer.org/skin-cancer-information/). The majority of these cases are caused by over-exposure to the sun's ultraviolet light and UV rays associated with tanning beds. Both light sources damage DNA leading to defective DNA repair mechanisms and cancers.

Breast Cancer

Cancer is typically named after the body part it affects, so breast cancer is the uncontrollable growth of cells in the breast tissue (9). Breast cancer in women accounts for 1 in 10 newly diagnosed types of cancer each year in the United States and is one of deadliest forms of cancer for women in the world (10). Breast cancer can develop silently for years. Most breast cancers are detected during routine screenings while others are discovered by the patient after finding lumps in the breast (10). The anatomy of the breast (Figure 2) consists of milk producing glands that lie on the pectoralis major muscle behind the chest wall (11). A collection of approximately 15-20 lobes forms in a circular arrangement, while fat surrounds the lobes determining the size and shape of the breast (10). The lobes are interconnected by thin tubes called breast ducts, which carry milk produced by the lobes to the nipple (12). Most breast cancers tend to develop in the ducts of the breast tissue while others begin in the lobules and other tissues of the breast (9). The metastatic form of breast cancer can penetrate the boundaries of breast tissue, migrate and develop in other tissues of the body including the lungs, brain, bones and liver (12). Early detection of breast cancer greatly improves the prognosis of the disease. Mammography and other screening techniques like Magnetic Resonance Imaging (MRI) are proving to be affective in the detection of breast cancer, increasing patient survival rate by over 80% (12). Alternatively, breast cancer statistics are drastically different in underdeveloped countries. Underdeveloped countries lack the resources and infrastructure needed for early detection of breast cancer, as a result, early detection rates are lower and mortality rates are higher compared to that in developed countries (13). According to The World Health

Organization (WHO), early detection of breast cancer greatly improves breast cancer outcome and survival and is the foundation of regulating the disease (13).

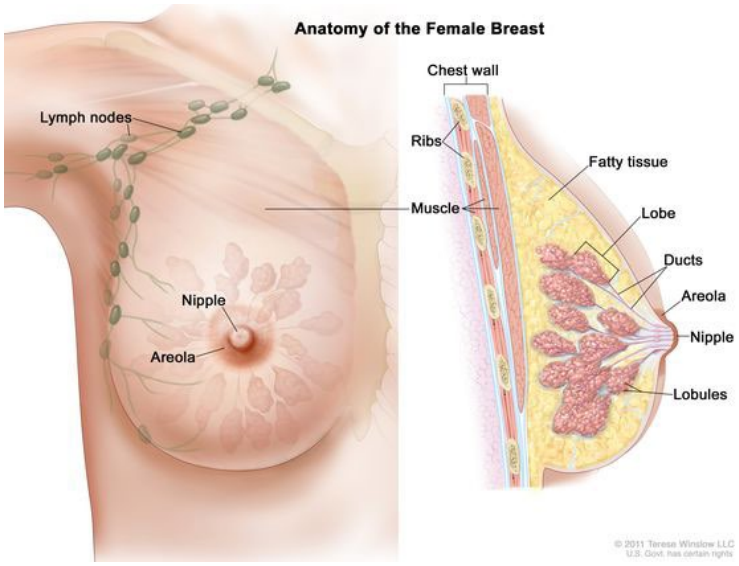


Figure 2: Diagram Illustrating the Anatomy of the Female Breast

Breast Cancer Treatment

Breast cancer treatments typically depend on the type and stage of the cancer. Modern medicines and medical therapies have been employed to prevent and treat breast cancer growth and development (14). Some of the most common forms of breast cancer treatment include surgery, chemotherapy, hormonal therapy, targeted biological therapy, and radiation therapy. Often, people with breast cancer undergo one or more kinds of treatments, and different breast cancer specialists work together to treat the patient (14). The leading approach to treating breast cancer today involves 'breast-conservation surgery' (to remove the cancer and the immediate surrounding areas) followed by an adjuvant therapy (i.e., treatment after the primary treatment) to decrease the risk of secondary

malignant growths and ensure a full recovery (15). Radiation and chemotherapy treatments are provided after surgery, to destroy any cancer cells not observed during the surgical procedure; this reduces the risk of a localized reemergence of cancer (15). Radiation therapy involves exposing the cancer cells to high levels of direct radiation to shrink the tumor. Side effects to radiation therapy include decreased sensitivity in the breast tissue along with cosmetic and other problems in the treated areas of the tumor (15). Conventional targeted therapy (like tamoxifen) is used to effectively treat estrogen and progesterone positive patients, and trastuzamab is used to treat HER2 positive patients.

Treatment of TNBC

Because cells associated with TNBC cancers are receptor-negative, the targeted therapies noted above are not effective and cannot be utilized for treatment of TNBC. Radiation and chemotherapy are most often used for treatment of TNBC patients. Current neoadjuvant chemotherapy has proven efficacy in the treatment of early-stage and advanced TNBC.

A review published by Lehmann et al. titled "*Clinical Implications of Molecular Heterogeneity in Triple Negative Breast Cancer"* in 2015 set out to review the molecular heterogeneity of TNBC and how this diversity impacted previous and future clinical trials (16). TNBCs lack ER, PR and HER2 gene expression making them insensitive to anti-hormonal and HER2 targeted therapies (16). Lehmann found that about 30% of patients diagnosed with TNBC benefit from neoadjuvant chemotherapy and those that experienced a pathological complete response during surgery showed significant improvements in overall survival (16). Typically, patients with TNBC have much lower survival rates than other breast cancer diagnosis even though they show a better response to chemotherapy

(16). This is likely due to chemotherapy-resistant tumors that remain after treatment for many TNBC patients (16). Results from these and other studies by Lehmann et al studies support the utility of examining precise genetic differences among TNBC patients so that a more 'targeted treatment strategy' can be use. This approach can only be applied if the genetics of TNBC patients are more accurately characterized, which the Lehmann laboratory has accomplished. Much of my research is based on analyzing TNBC using the Lehmann classification, 'instead of analyzing the cancers as a single homogenous group'. My aim is to (a) identify genetic differences between African American (AA) and Caucasian (CA) TNBC that will be studied for their potential role in contributing to cancer in AA patients. Ultimately, our aspiration is that genes discovered in our laboratory will lead to strategies that impact patient survival.

Breast Cancer Types

Breast cancer is not just one type of cancer. There are different forms of the cancer which are named depending on the original location of the cancer and its progenitor cell type (17). The type of breast cancer will ultimately determine the type of treatment or therapy. Breast cancers can be defined based on their pathology and more recently with the advent of molecular analysis tools (like DNA microarrays), breast cancers are defined based on molecular descriptions.

Breast Cancer Types Based on Pathological Diagnoses

The pathological distinctions of breast cancers are summarized below. Molecular characterizations of breast cancers are described later.

- *Lobular Carcinoma in Situ (LCIS)*: defined as an abnormal, in situ cell growth in the lobules.
- *Ductal Carcinoma in Situ (DCIS):* non-invasive, in situ, originates in the milk duct; incidence up to 20% of cancers diagnosed.
- *Invasive Ductal Carcinoma (IDC):* invasive, infiltrating ductal cancers, occurring in up to 80% of cancer diagnosed.
- *Tubular Carcinoma of the Breast:* a subtype of the IDC cancers. The cells have a tubular configuration, with an incidence of 8%-27% in patients.
- *Medullary Carcinoma of the Breast*: rare invasive ductal carcinomas with resemblance to the brain's medulla. They occur in ~3% of patients.
- *Mucinous Carcinoma of the Breast:* rare cancers described as 'floating in a pool of mucin'. The cancer begins in the milk ducts.
- *Papillary Carcinoma of the Breast*: rare invasive carcinomas with finger-like projections, occurring in approximately 1-2% of patients.
- *Cribriform Carcinoma of the Breast*: rare, normal-looking invasive cancers named based on their cribriform-like configurations.
- *Invasive Lobular Carcinoma (ILC):* invasive cancers that originate in the lobules of the breast. They occur in ~20% of patients.
- *Inflammatory Breast Cancer (IBC):* rare very aggressive, invasive cancer that spread within the breast tissue in 'sheets' instead of lumps, making the cancers difficult to detect and ultimately treat.

- *Phyllodes Tumors of the Breast*: rare tumors, occurring in less than 1% of patients. The cells grow in a leaf-like configuration.
- *Paget's Disease of the Nipple*: rare breast cancer involving cancer around the nipple and ducts draining towards the nipple.
- *Male Breast Cancer:* Breast cancers in men are rare, occurring at a rate of less than 1%. Breast cancers in males are smaller than those observed in females, but they can be equally as invasive.

Breast Cancer Types Based on Molecular Description

The first DNA microarray was proposed by Patrick Brown in early 1980's (ref), and manufactured as large scale in the 1990's. Much of the data related to the molecular description and signaling pathways associated with breast and other cancers are identified using the microarray platforms. Microarrays allow for more detailed biological characterization of the genome and transcriptome of cellular mechanisms, which in turn allow for more accurate diagnoses and subsequent treatment (18). The molecular descriptions of breast cancers are summarized below. There are six molecular subtypes of breast cancer which include (Figure 3):

- *Luminal A*- cancers are defined as positive for estrogen receptor and progesterone and negative for HER2 gene expression and low Ki-67 levels.
- *Luminal B*- cancers are defined as positive for estrogen receptor and progesterone receptor and for positive for HER2 or negative for HER2 with high Ki-67 levels.

- *HER+ve* (HER2neu/ErbB2)- cancers are enriched in HER2 levels and negative for estrogen and progesterone receptor expression.
- *Normal-like-cancers* are similar to Luminal A but follow normal breast profiling.
- *Triple negative*- negative for estrogen, progesterone and HER2 expression. Expanded description in the section below.
- *Claudin low*-cancers display low claudin (cell adhesion) expression and high epithelial-mesenchymal transition genes.

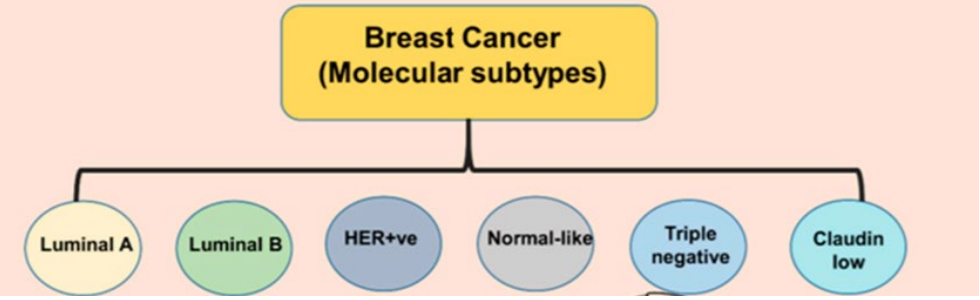


Figure 3: Molecular Subtypes of Breast Cancer
Triple negative breast cancer is one of the six types of breast cancer based on molecular characterization (6).

TNBC (triple negative breast cancer) are negative for the three receptor genes estrogen receptor, progesterone negative and HER2neu. Of the 6 molecular subtypes, TNBC patients represent ~15% of breast cancers and have low 5-year survival rates. Although TNBC are characterized as negative for 3 receptor genes, Lehmann et al show the subtype can be further divided into six sub-categories based on clustering signatures and molecular gene expression (19). DNA microarrays were utilized to further characterize TNBC and identify the various sub-categories. Lehmann et al. identified: (a) two basal-like types (designated BL1 and BL2)- basal type 1 (BL1) includes over-expression of genes involved in signaling pathways related to Ribonucleic acid (RNA) polymerase processes, cell division and cell cycle regulation. Patient samples characterized as Basal type 2 (BL2) genotype are thought to be of myoepithelial origin, including genes associated with growth factor signaling processes, gluconeogenesis, and glycolysis. (b) luminal androgen receptor or molecular apocrine group (LAR)- this group includes estrogen receptor samples with over-expression of androgen receptor, and other genes involved in hormonal regulation. (c) an immunomodulatory group- samples characterized in the immunomodulatory (IM) category appear like medullary breast cancers and are enriched in genes involved in immune signaling pathways, cytokine signaling, natural killer cell pathway and antigen identification and processing. (d) a mesenchymal (M) and (e) mesenchymal stem-like group (MSL) the group- The M and MSL groups are enriched in genes associated with cell motility, mesenchymal-like differentiation, proliferation, and extra-cellular matrix proteins. There are significant implications to these studies. Lehmann et al suggested that ‘by defining TNBC based on unique signatures, driver signaling pathways can be identified

and used as pharmacological targets. These transcriptome distinctions are what I refer to as Lehmann classification in this text.

Anatomical Location of TNBC Progenitor Cells in Breast Tissue

Breast cancer can be characterized by genetics, and by pathology. A normal microscopic representation of breast tissue is given in Figure 4 (20). Data suggest the progenitor cells for TNBC line the outer regions of the luminal regions. These cells are described as basal-like cells, which surround the myoepithelial and luminal cells, each which are closer to the breast lumen.

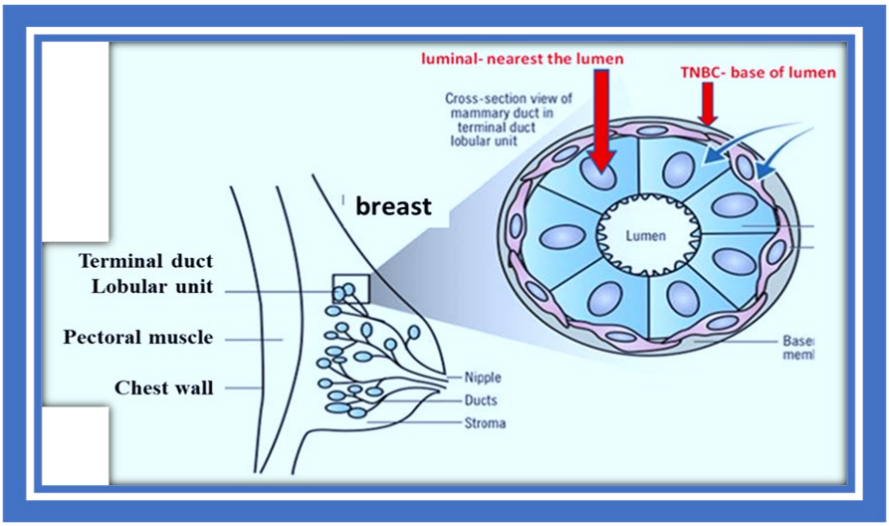


Figure 4: Diagram of the Breast Tissue and Luminal Ducts
Anatomical location of Triple Negative Breast Cancer can be found in the basement membrane of the luminal ducts (7).

Triple Negative Breast Cancer and Racial Disparities

TNBC are based on molecular characterizations. The primary focus of my research is to further characterize the cancers. We have several experimental approaches aimed at understanding the cancers. Data show racial disparity in AA compared to CA patients. We

will examine the transcriptome of AA and CA samples, with the goal of identifying genes differentially expressed between the patient populations.

Metadata analyses show that racial disparity is due to three principal determinants (Figure 5): (a) the difference in obesity among the two ethnic groups (b) socioeconomic factors such as under representation in clinical trials or simply not having access to suitable hospital resources and (c) biological determinates related to race. The biological determinants include differential mutations associated with BRCA1 and/or p53 genetic factors, and differential expression of cancer stem cell genes like ALDH1 and Wnt-signaling pathway genes (21). This research also focused on interrogation of samples isolated from AA and CA patients in search of defining biological differences between the two ethnic groups.

There is a higher incidence of TNBC in AA and women of African descent compared to CA and women of European descent. In 2011, a study by Lehman and others showed that TNBC was not just one type of cancer in patients, but a complex heterogenous type of breast cancer that could be subdivided into at least 6 different subdivisions (19). In addition to addressing the complexity defined by Lehmann, we are addressing the biological differences between AA and CA TNBC genotypes based on Lehmanns' classification.

A disproportionate number of the earlier TNBC studies were performed analyzing CA patient samples. So, we have a lot to learn related to the biological processes in AA TNBC. Since much of what is known about TNBC has been obtained following microarray analyses, we will utilize DNA microarray to further compare the ethnic populations. We will characterize TNBC in CA compared to AA patients, focusing on separate analysis of

Basal A and Basal B subgroups that define the TNBC with the aim of identifying gene expression patterns that contribute to the disparity observed between the two races. Identifying the differences between the two ethnic populations could lead to a better understanding of TNBC tumors in general as well as the development of biomarkers and therapeutic targets to improve patient survival. Our hypothesis is that there are genetic signatures that distinguish TNBC in CA compared to AA patients which can be identified by comparing sub-populations within the TNBC subtype. Our analysis can be summed up into one principal approach. We aim to identify differentially expressed genes related to either Basal A or Basal B subtypes associated with CA and AA cell lines.

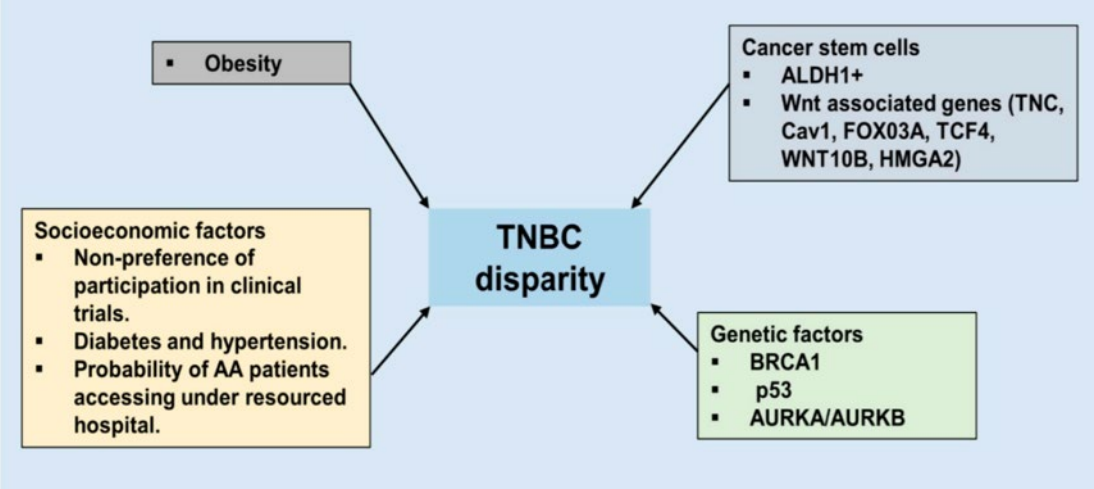


Figure 5. Racial Disparity

TNBC in women of African-descent vs Caucasian based on Obesity, Socioeconomic determinants, and Biology (genes) (25).

CHAPTER 2

LITERARY REVIEW

An incredible number of studies show that AA patients have a higher incidence of TNBC and age-adjusted mortality. AA patients are twice as likely to be diagnosed with the disease and more likely to die from breast cancer. Many investigators suggest that biological factors and not just socioeconomic factors contribute to these disparities (22). Data confirms that the microarray experimental platform is the best experimental approach to use for this type of study. These analyses are also best performed using approaches that consider the genetic heterogeneity and complex genetic signatures of the TNBC. As a result, comparative analyses of CA versus AA patient must consider the heterogeneity of TNBCs as defined by Lehmann et al. Our analyses are based on both microarray and the Lehmann approach. Once TNBC are defined based on genetic heterogeneity, then more reliable signatures distinguishing CA vs AA can be found and considered as therapeutic strategies. We must emphasize that other laboratories are aware that this is the most logical approach towards better characterizing differences based on ethnicity.

Significance of Lehmann et al. Studies

Even though TNBC is defined as negative for 3 receptors it is a molecularly diverse disease (23). This diversity limits the success of traditional targeted therapies in patients (23). Studies by Lehmann et al have been instrumental in further characterizing the

cancers. Lehmann used DNA microarray bioinformatic analyses to further divide TNBC into six sub-categories defined by clustering of molecular gene signatures. This unique characterization of the transcriptional subtypes of TNBC has led to a more 'focused use of conventional therapies and identification of *potential* biomarkers being studied as targeted therapies for TNBC patients (24). New biomarkers (identified using the Lehmann approach) are currently being studies for their use as therapies for TNBC patients (23).

Genomic and Socioeconomic Relationship of TNBC Among AA vs CA

Data are clear in that disparities in AA compared to CA are multifactorial, complex and related to both socioeconomic and biological factors. The socioeconomic factors are related to AA's predisposition to diseases like diabetes and hypertension, somewhat higher levels of obesity and health care disparities based on access to healthcare like mammographies, delivery of treatment and psychological and cultural factors. These factors have been studied in recent years and found to enhance or inhibit the outcome of the biological factors (25). In a 2013 study by Danforth entitled *"Disparities in breast cancer outcomes between Caucasian and African American women: A model for describing the relationship of biological and nonbiological factors",* he suggested that many of the nonbiological disparities between AA women and CA women could be modified. Once modified, the nonbiological factors could drastically improve the outcome of the biological factors (25). The biological factors are thought to be mostly related to differences in the incidence of BRCA1 and p53 mutations, variations in Wnt signaling pathways and other cancer stem cell signaling irregularities (ref).

The socioeconomic and genetic factor disparities of TNBC are widespread and affect women of African descent not just African American women. One study showed

that TNBC was the most dominant cancer diagnosis in sub-Saharan Africa and in 22 countries throughout the America's and Caribbean (Figure 6) (22). The higher risk of TNBC in African American women and sub-Saharan women versus CA and European women further suggest that specific genetic components of geographically defined African heritage are associated with hereditary susceptibility of TNBC carcinogenesis (26).

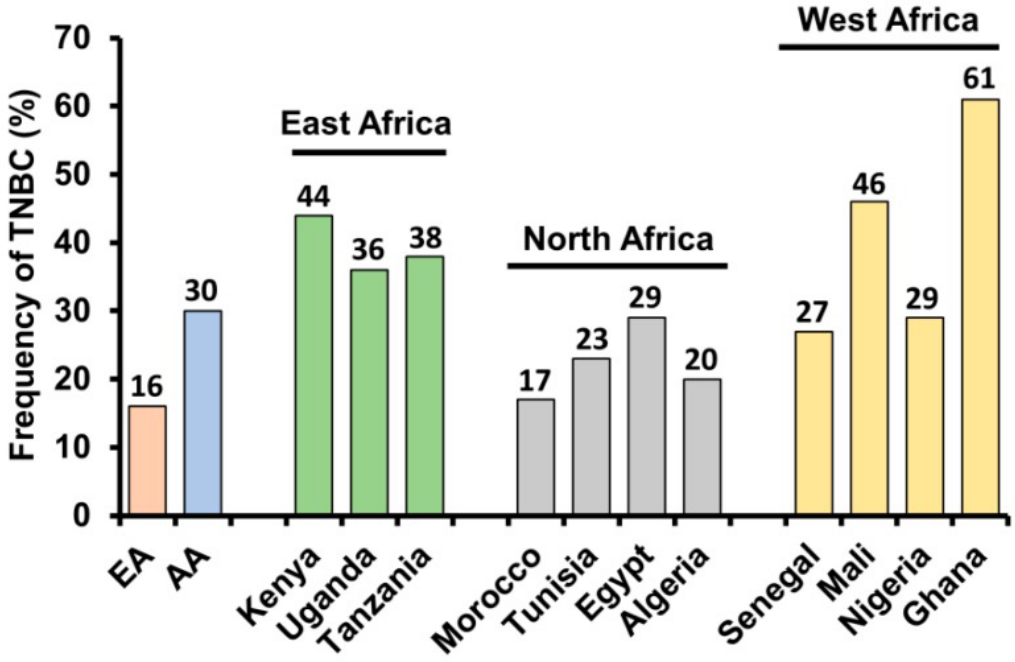


Figure 6: TNBC Frequency Among Women of African Descent vs Women of European Descent

<u>Continued Review of Gene-related Factors Related to Racial Disparities</u>

At this point, many of the studies demonstrating differences in dysregulation, differential gene expression and genomic mutations in AA compared to CA TNBC patients are merely observations, based on large scale analyses of archival datasets. In one such study p53 mutations were observed in >45% of AA compared to 27% of CA, MLL3 was

observed in 12% of AA compared to 6% of CA, and PIK3CA mutations in 23% of AA compared to 34% of CA (Ademuyiwa et al.). The p53 gene is the most frequent mutation found in cancer, MLL3 is a tumor suppressor gene deleted in myeloid leukemia, and PIK3CA is associated with the PI3K pathway which is also frequently altered in cancers. In a separate study Lee et al. (27) identified the gene resistin overexpressed in AA. Resistin is secreted by adipocytes and is suspected of linking obesity to type II diabetes leading to an intriguing hypothesis connecting obesity to a molecular signature in AA patients. In yet another study, Field (28) matched tumors based on pathological characteristics and observed CRYBB2, PSPHL, and SOS1 genes differentially expressed in AA compared to CA patients. The three genes are shown to be involved in cellular growth and differentiation, metastasis and invasion and immune response: another intriguing study correlating the biological characteristics of TNBC in AA patients with functional pathways and genes.

Epidemiological Risk of TNBC in AA Women

Many epidemiological risk factor differences have been identified as associated with TNBC in AA women vs CA women. It should be mentioned that these are correlative studies. Reproductive factors like menstruation beginning at younger ages and younger ages of full-term pregnancy, higher exposures to societal inequality, and shorter duration or a complete lack of breast feeding are thought to contribute to epidemiological differences in TNBC disparities. In addition, the size and shape of the body like higher body mass index and waist to hip ratio correlate with the epidemiological differences in TNBC disparities (29). Sturtz and colleagues evaluated the differences in epidemiological factors and gene expression profiling between AA women and CA women (29). This

evaluation was referred to as the Clinical Breast Care Project. The investigators examined the incidence of obesity, estrogen exposure, breastfeeding, diet and physical activity and co-morbidities with TNBC status (29). They found that of the 1064 AA women and CA women evaluated 15% of the women had TNBC. Of the women with TNBC, the incidence in AA patients was 28%, compared to an incidence of 12% in CA patients (29). Furthermore, the frequency of TNBC was higher in premenopausal AA women (53%) compared to CA women (42%) (29). They also found that AA women were more likely to be obese and that caffeine and alcohol use was significantly lower in AA women (29). Another study performed Trivers et al sought to evaluate whether anthropometrics (i.e., science related to physical size and form), demographics and reproductive history were associated with distinct breast cancer subtypes (30). They analyzed 460 women, 116 black women and 360 white women, all with different subtypes of invasive breast cancer (30). The different types of breast cancers were grouped as TNBC (ER-,PR-, HER2-), (ER-,PR-HER2+), (ER/PR+, HER2+), (ER/PR+, HER2-) (30). They found that women with TN tumors were more likely to be obese and that no matter the HER2neu status ER-PR- tumors were more likely associated with AA women, younger age at first birth, having a recent birth and being overweight (30). Collectively, these data suggest both socioeconomic and biological factors contribute to disparities between AA compared to CA TNBC patients. The multifactorial nature of the disease makes discovery of a ‘single cause’ a daunting task.

CHAPTER 3

DESIGN OF THE STUDY

Summary Statement

Data examined in this thesis are based on the bioinformatic analyses of subsets of AA compared to CA TNBC cell lines and patient samples previously processed using DNA microarray platforms. All datasets were extracted from Gene Expression Omnibus (GEO) associated with the National Center for Biotechnology of the National Institutes of Health.

Rationale for Selection of the Cell Lines Used in this Study

Prior to the experimental results published by Lehmann et al. (19), practically all experimental analyses of TNBC were based on treating TNBC as a single unit with limited regards to the cancer's heterogeneity and biological complexity. Lehmann wasn't the first, but his studies were one of first to elegantly demonstrate that the single TNBC subtype could be further characterized based on clustering of precise genetic signatures. Further characterization of TNBC can lead to more accurate results when comparing TNBC patients to each other, and when comparing TNBC patient transcriptomes to other breast cancer subtypes. Our (and other's) approach to analyses of AA to CA patients are based on considering TNBC sub-categories as defined by Lehmann et al.

We were aware that a limited number of AA TNBC cell lines were available for this study. So, we first identified TNBC cell lines isolated from AA patients (as defined by

references documented at atcc.org and GEO) and grouped them according to Lehmann's designation. To allow for accurate comparisons, we identified CA TNBCs that closely resembled the AA based on molecular signatures (Table 1).

Table 1: Lehmann Classification and Genetic Characterization of TNBC

TNBC subtype	Genetic abnormalities	Cell line	Subtype correlation[A] (*p* value)	Histology	Mutations
Basal-like 1	Cell cycle gene expression DNA repair gene (ATR-BRCA pathway) Proliferation genes	HCC2157 HCC1599 HCC1937 HCC1143 HCC3153 MDA-MB-468 HCC38	0.66 (0.00) 0.62 (0.00) 0.28 (0.00) 0.26 (0.00) 0.24 (0.00) 0.19 (0.06) 0.19 (0.01)	DC DC DC IDC DC DC	BRCA1; STAT4; UTX BRCA2; TP53; CTNND1; TOP2B; CAMK1G BRCA1; TP53; MAPK13; MDC1 TP53 BRCA1 PTEN; RB1; SMAD4; TP53 CDKN2A; TP53
Basal-like 2	Growth factor-signaling pathways (EGFR, MET, NGF, Wnt/β-catenin, IGF-1R) Glycolysis, gluconeogenesis Expression of myoepithelial markers	SUM149PT CAL-851 HCC70 HCC1806 HDQ-P1	0.30 (0.00) 0.25 (0.00) 0.24 (0.04) 0.22 (0.00) 0.18 (0.17)	INF IGA DC ASCC IDC	BRCA1 RB1; TP53 PTEN; TP53 CDKN2A; TP53; UTX TP53
Immunomodulatory	Immune cell processes (CTLA4, IL2, IL7 pathways, antigen processing/ presentation) Gene signature for medullary BC (rare TNBC with a favorable prognosis)	HCC1187 DU4475	0.22 (0.00) 0.17 (0.00)	DC DC	TP53; CTNNA1; DDX18; HUWE1; NFKBIA APC; BRAF; MAP 2 K4; RB1
Mesenchymal-like	Cell motility Cell differentiation Growth factor signaling EMT	BT-549 CAL-51 CAL-120	0.21 (0.00) 0.17 (0.00) 0.09 (0.00)	IDC AC AC	PTEN; RB1; TP53 PIK3CA TP53
Mesenchymal stem-like	Similar to M+ Low proliferation Angiogenesis genes	HS578T MDA-MB-157 SUM159PT MDA-MB-436 MDA-MB-231	0.29 (0.00) 0.25 (0.00) 0.14 (0.00) 0.13 (0.00) 0.12 (0.00)	CS MBC ANC IDC IDC	CDKN2A; HRAS; TP53 NF1; TP53 PIK3CA; TP53 HRAS BRCA1; TP53 BRAF; CDKN2A; KRAS; NF2; TP53; PDGFRA
Luminal androgen receptor	Androgen receptor gene Luminal gene expression pattern Molecular apocrine subtype	MDA-MB-453 SUM185PE HCC2185 CAL-148 MFM-223	0.53 (0.00) 0.39 (0.00) 0.34 (0.00) 0.32(0.00) 0.31 (0.00)	AC DC AC AC	PIK3CA; CDH1; PTEN PIK3CA PIK3CA PIK3CA; RB1; TP53; PTEN PIK3CA; TP53

Data from Lehmann et al. [14]

On the far right is the genetic characterization of each of these subtypes. In the center of the diagram, we have cell lines that have been isolated from patients and associated with their respective subtypes. The cell lines that we examined for our study are designated by the arrows. The African American cell lines are designated by the arrows with the red boxes and the Caucasian cell lines are designated with just an arrow. The top four arrows are the cell lines that are being compared for the Basal A group and the bottom three arrows are the cell lines of the Basal B group.

Table 1 details the 6 sub-categories of TNBC as defined by Lehmann (19). On the far left of the table is the Lehmann classification of the types of TNBC. All cell lines in the

Table 1 are TNBC and have been characterized based on gene analyses. Each of the sub-categories are characterized based on Genetic abnormalities and Mutations. The particular cell line associated with their respective subtype is noted in the center column of the Table. The cell lines that we examined for our study are designated by the arrows. The AA cell lines are designated by the arrows with the red boxes and the CA cell lines are designated with just an arrow. The ethnicities of the cell lines were documented by references available at attc.org and GEO. These cell lines were grouped according to their genetic similarities based on Lehman classification, so that we could compare genetically similar AA to CA patient cell lines. Using this approach, we can infer that after our analyses, to some degree, differential gene expression is related to race. Cell lines in the top portion of the table are designated Basal A group; four cell lines are included in this group. Cell lines in the bottom portion of the table are designated Basal B; three cell lines are included in this group. It's important to note that there are limitations of my research; there are a limited number of AA TNBC cell lines available for study. The only well-defined AA cell lines are the ones designated here. The summary of the AA vs CA cell lines used in our study are given in Table 2. The DNA microarray experimental platform was utilized to interrogate and compare transcriptomes between AA vs CA cell lines. Raw transcriptome datasets for the cell lines were obtained from GEO. The precise GEO raw data file for each cell line is given in the DNA microarray Section below.

Table 2 is a summary of the TNBC cell lines used for the Basal A comparison and the Basal B comparison in the study. Basal A cell lines are defined as basal-like and immunomodulatory characteristics. Basal B cell lines are defined by their mesenchymal / stem-like characteristics.

Table 2: **Summary of the TNBC Cell Lines**

• **Basal A:** two groups to separately compare (ie, group 1 vs group2)
• Group 1 (CA cell lines)-included HCC1143 and HCC1187
• Group 2 (AA cell lines)-included MDA MB468 and HCC70
• **Basal B**: two groups to compare (ie, group 1 vs group 2)
• Group 1 (CA cell lines)-included MDA MB231 and MDA MB436
• Group 2 (AA cell lines)-included MDA MB157)

DNA Microarray Analyses

GEO Source for the DNA microarrays for each cell line (raw data (not normalized). The 'wet-lab' microarray experiments were not performed by our laboratory. The microarray dataset for each cell line was obtained from GEO. Once in GEO, we used the search terms 'microarray and triple negative' to identify cell lines processed using microarrays. After we identified TNBCs processed using the microarray platform, we searched individual datasets for AA and CA cell lines. The original GEO dataset information is given below (in the event investigators wish to locate and download the files). Duplicate sources were identified for some cell lines. **Basal A** cell line GEO datasets: GSM1589133 (HCC1143), GSM276023 (HCC1143), GSM1589134 (HCC1187), GSM1589146 (MDA MB468), GSM276009 (MDA MB468), GSM158914 (HCC70), GSM276013 (HCC70). **Basal B** cell line GEO datasets: GSM1589131 (MDA MB436), GSM1589153 (MDA MB231), GSM275993 (MDA MB231), GSM1589152 (MDA MB157). Summaries of which cell lines are grouped together for comparison are outlined in Table 2 above. Individual cell lines used for the Cluster analyses were obtained from the GSE12777 dataset obtainable at GEO.

Microarray Procedural Outline: As summary, the microarray platform can be used to determine the transcript levels of RNA preparations (i.e., cell line preparations). The technique is used to interrogate the expression levels of a large number of genes at the same time, allowing for comparative transcriptome and signaling analyses between different samples. DNA microarrays (sometimes called gene-chips) are cassettes that consist of printed spots containing known DNA sequences. The DNA sequence corresponds to specific regions of a gene that can be used to probe for detection of that gene (25).

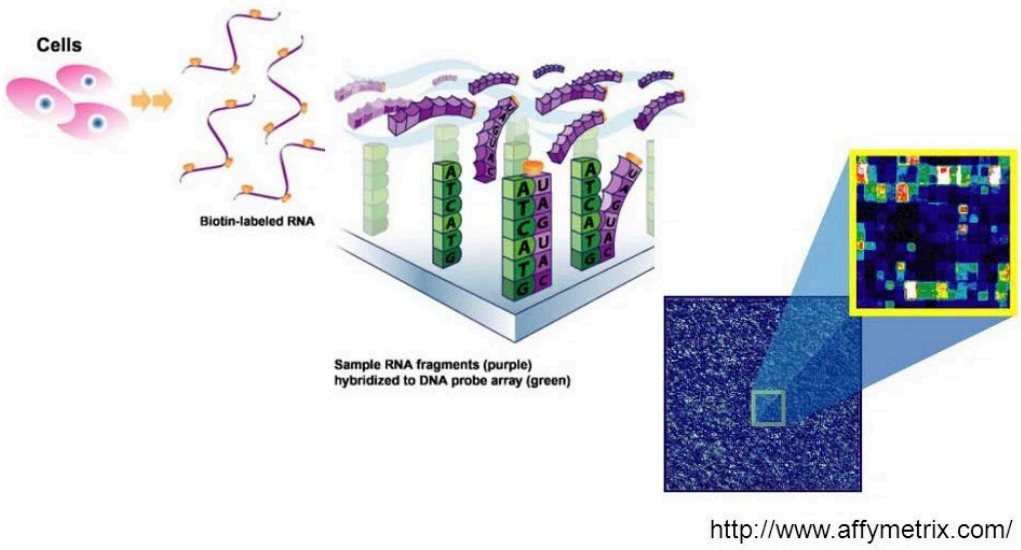


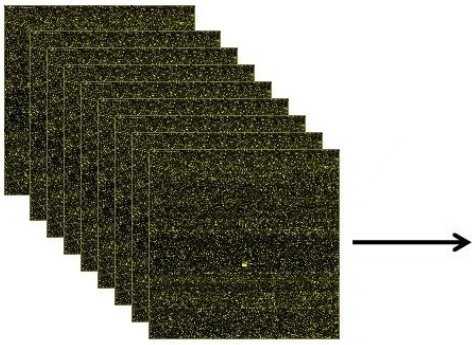
Figure 7: Diagram of a Gene Expression Array Illustrating How Biotin-labeled RNA Fragments (purple) Hybridized to DNA Probe Array (green) (29).

RNA is extracted from cells during log phase. mRNA transcripts are converted to antisense, end-labeled and hybridized to their complement sequence which is immobilized to a high-density gene-chip cassette. Probe-sets (i.e., sequences on the cassette) are positioned within microns of each other. For this particular microarray, the color ranges

from black which corresponds to zero or low levels of transcript detected to white, which corresponds to high copy number transcript; see the yellow insert.

The sense form of the gene is spotted and immobilized onto the cassette, which can be used to hybridize and detect the corresponding fluorescently labeled antisense complement (Figure 7). The amount of hybridization and subsequent florescence can be used to determine copy-number (based on comparisons to internal controls). The Affymetrix microarray platform is the most widely used microarray platform for full genome expression analysis and is the platform employed in my project (25).

.



	sample 1	sample 2	sample 3	sample 4	.....
gene 1	9.249132	9.771213	9.390076	9.395176	.....
gene 2	6.989496	5.84592	6.063214	4.995175	.....
gene 3	4.549009	5.298832	4.028992	4.730776	.....
gene 4	7.042218	7.156791	6.516016	6.4736	.....
gene 5	2.842815	3.210668	3.168886	3.203355	.....
gene 6	6.076624	6.255116	5.53142	7.186467	.....
gene 7	4.001927	4.408226	4.426111	4.218325	.....
gene 8	4.011074	4.147679	3.506027	3.450706	.....
gene 9	6.374999	7.199643	5.660234	8.143042	.....
gene 10	2.905368	3.040598	2.994988	3.029243	.....
gene 11	3.710801	3.787264	3.713254	3.393635	.....
gene 12	6.687304	7.583114	7.201481	7.436545	.....
gene 13	3.64797	3.731533	3.634993	3.800914	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
gene 50000	7.40968	7.862199	7.34268	6.801629	.....

Figure 8: Example of Gene Expression Data Generated from Microarray Gene Chips (29)

Approximately 50,000 genes are compared across four microarrays. Transcript levels are displayed as log2 values. A pseudo-color can be used to correspond to gene transcript copy number. The colors differ depending on the user preference. Colors in this figure range from black which represents transcript levels near zero, to bright green which corresponds to high transcript copy number.

One microarray cassette or gene-chip is used for each sample. After the sample is hybridized to the array and analyzed by the software, the data is normalized, compared to internal controls and transcript levels for each sample are displayed (row-by-row) allowing the genes on different microarrays to be directly compared (Figure 8). The values given in Figure 7 represent log 2 values corresponding to the transcript levels for 4 different samples, for ~50,000 gene transcripts. For my experiments, we downloaded cell lines processed using the Affymetrix U133 plus 2.0 microarrays which contained 54,676 transcripts and transcript variants.

Data Analyses

MAdB Online Database

mAdB is a private online data analyses resource supported by NCI. mAdB contains data analyses tools suitable for miRNA, genomic, transcriptome, pathway, and other types of bioinformatic tools (mAdb.nci.nih.gov). The TNBC cell line microarray datasets were uploaded from GEO to mAdB. I used mAdB to determine which genes from Basal A cell lines (CA-HCC1143, CA-HCC1187, AA-MB468, AA-HCC70) and from Basal B cell lines (CA-MDA-MB231, CA-MDA-MB436, AA-MDA-MB157) possessed the greatest degree of transcript copy number difference in expression level (i.e., differential expression) (30). Refer again to Table 2 for a summary of the comparisons. To compare Basal A samples, CA was designated as group 1 and AA was designated as group 2 and T-test with unequal variance was applied. Separately to compare Basal B samples, CA was designated as group 1 and AA was designated as group 2 and T-test with unequal variance was applied (Figure 9, Figure 10).

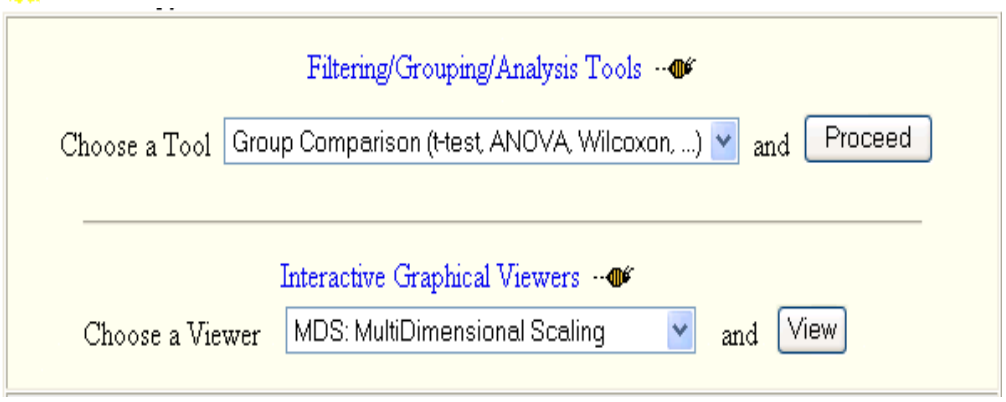


Figure 9: Snapshot of the mAdb Two-group Test Program Interface (30)

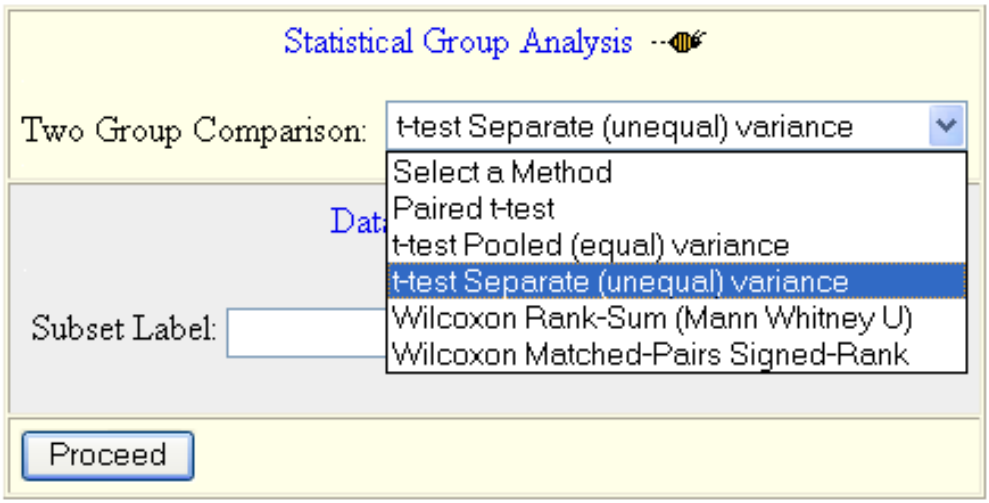


Figure 10: Snapshot of the mAdb 2 Group Statistic Analysis Automatically Selected for a 2-group Dataset (30)

T-test Group Analyses

To identify differentially expressed genes (i.e., a list of genes with expression levels statistically and (more important) biologically different in two or more sets of the representative transcriptomes) a T-test group analysis was done. The t-test group analyses determine whether the means of two groups are statistically different. Following t-test

analyses, gene are selected for further analyses if there is at least a 2-fold difference between transcript levels, with p-value <0.05; this is the industry accepted criteria for reliable differentially expressed genes using the microarray platform.

Conserved Domain Analyses

The National Center for Biotechnology Conserved Domain (CD) Blast search tool (31) was used to perform CD analyses. The CD analyses was utilized to identify conserved sequence motifs present in differentially expressed genes. CD similarities imply functional similarity between different genes and proteins.

Signaling Pathway Analyses

Signaling pathway analyses were performed using mAdB.

STRING protein: protein interaction Analysis-

Retrieval of Interacting Genes/Proteins (STRING) is used to determine the relationship between proteins based on millions of data points related to experimental data, computational prediction methods and public text collections (32).

Hierarchical Cluster Analysis of TCEALs:

The hierarchical (HC) analyses were used to determine genes similarly expressed. This was performed using mAdB.

Multalin Analysis of TCEALs

Multalin is an online sequence analysis program. Multalin sequence analysis was used to validate the CD of differentially expressed genes (32). Multalin was performed by submitting the gene sequences to the online server.

CHAPTER 4
RESULTS AND DISCUSSION

As mentioned, two separate analyses were performed, AA vs CA based on the Basal A molecular distinction and AA vs CA based on the Basal B distinction. The data were generated using the DNA microarray platform which allows for comparative analyses of gene transcript levels. An example of these data is presented in Tables 3 and 4. The T-test was used to identify differences between the ethnicities. The T-test analysis compares the average gene expression levels between the two patient populations. The comparison values are plotted under the A-B mean difference column. T-test generates the fold change and the significance (of this change) in the form of the p-value. The data can then be sorted based on fold change and p-value. The industry standard for selecting differentially expressed genes in microarray is p-value of <0.05, with a fold change of >2.0. Even though the microarray initially contained 54,000 genes, most of the genes were not different between the populations; so, filtering based on p-value and fold change led to a substantially smaller list of genes. These genes were considered differentially expressed. As example of how the comparisons are presented after processing, see Table 3 and Table 4. After a shorter gene list is generated the tables can be transported into Excel, so additional sorting can be when performed. Cell line designations are included in the headings.

Table 3: **T-test Basal A Group**
Analysis comparing CA cell lines (1-3) to AA cell lines (4-7). Blue are CA and red are AA cell lines.

GSM1589133 HCC1143 maire	GSM1589134 HCC1187 maire	GSM276023 HCC1143 GENE COM	GSM1589146 MDA468 MAIRE	GSM1589149 HCC70 MAIRE	GSM276009 MDA468 GENE COM	GSM276013 HCC70 GENE COM				
#1	#2	#3	#4	#5	#6	#7	A-B p-Value	A-B Mean Difference	Feature ID	Gene
79.1	181.9	227.7	717.3	628.3	567.6	663.6	0.042304	-2.1118	1554878_a_at	ABCD3
417.4	178.5	95.8	1504.2	1197.7	524.3	1514.4	0.032764	-2.5056	201963_at	ACSL1
125.1	28.0	144.0	722.8	913.3	1163.4	1855.4	0.025434	-3.7794	238689_at	ADGRF1
811.4	1005.6	366.2	59.5	95.4	61.5	71.1	0.010986	3.2439	226325_at	ADSSL1

Table 4: **T-Test Basal B Group**
Analysis comparing CA cell lines to AA cell lines. Blue are CA (1-3) and red is the only AA cell line present (4-5). The program does not allow 1 sample, so the data were repeated.

GSM1589131 mda 436 basal b	GSM1589153 MDA231 MAIRE	GSM275993 MDA231 GENE COM	GSM1589152 basal B AA mda 157						
#1	#2	#3	#4	#5	A-B p-Value	A-B Mean Difference	Well ID	Feature ID	Gene
816.2	971.6	564.4	139.2	139.2	0.008716	2.4582	1493186	203504_s_at	ABCA1
779.8	1273.3	1050.7	217.5	217.5	0.0084732	2.2212	1473066	231907_at	ABL2
118.3	275.0	293.0	34.0	34.0	0.024611	2.6387	1494194	205730_s_at	ABLIM3
137.6	276.2	246.1	1044.2	1044.2	0.017664	-2.3093	1469761	227962_at	ACOX1

Table 3 and 4 are examples of the T-test group analyses data. The values in table (corresponding to each gene) represent the normalized transcript level for each gene. Only a portion of the table is displayed as an example. Table 3 shows an example of comparisons between the Basal A subcategory containing CA cell lines (in blue; 1-3) and the AA cell lines (4-7; red). Table 4 shows an example of the comparisons between the Basal B sub-category containing CA cell lines (in blue; 1-3) compared to the only AA cell available (in

red; 4-5) for the Basal B sub-category. For both tables, to the right of each area the significance p-value and the and the mean difference between each gene is noted by A-B p-value and A-B mean, respectively. The A-B mean represents the fold difference in transcript level for a particular gene. The gene symbol is in the final column. The comparisons were generated by --"CA divided by AA".

After the data were filtered (based on p-value and fold change), < 500 genes were generated. Built into mAdB is a function that will filter for background noise or questionable gene values. Using this program function, an even smaller, more manageable gene list is generated. Gene lists can be processed in Excel and online analyses tools like Molbiotools can be used to compare gene lists (http://www.molbiotools.com/listcompare.html).

Short List of Genes Common to Basal A and Basal B

There were a few genes common to both datasets, for example genes generated by comparing CA vs AA Basal A genes; and genes generated by comparing CA vs AA Basal B genes (Table 5). The short list of genes was identified using the online tool (http://www.molbiotools.com/listcompare.html). These genes were common to Basal A and Basal B. Most of the genes were not the same but likely belong to the same gene family. A table of the genes that were common across Basal A and Basal B along with their differential gene expression levels are described in Table 6. The values represent a comparison between CA vs AA, so if the value is positive that means the gene is downregulated in AA compared to CA, and if the value is negative the gene is high in AA. Ideally, we searched for the exact same gene on both Basal A and Basal B cell lines that

showed the same direction of differential expression (if down-regulated in Basal A then also down-regulated in Basal B). That would suggest "a sort of AA gene". Because we did not find this gene, it could be that (a) no such gene exists or (b) there are significant differences between Basal A and Basal B that '*different AA genes drive or, are associated with the signaling processes in the 2 sub-categories*. Even though CCDN1, CYP4X1 and SUSD2 genes satisfied this requirement, when we examined the genes further and compared them to independent datasets of CA cell lines (i.e., MDA MB231; Dr Player's datasets), the genes did not show the same pattern of expression, in other works, the results were not reproducible.

Table 5: Short-list from Basal A and Basal B
Analyzed to determine common genes between both cell lines. Using this online tool-Molbio-Tools (http://www.molbiotools.com/listcompare.html)

BASAL B genes	BASAL A genes
CCND1	CCND1
CLDN10	CLDN8
CYP4X1	CYP4X1
EEF7	E2F5
EIF4EBP1	EIF5A
HIST1H2BG	HIST1H2BD
HIST1H2BG	
SUSD2	SUSD2
TCEAL8	TCEAL9

Table 6: Table of Common Genes
Showing their direction of differential expression (If up-regulated in Basal A then up-regulated in Basal B).

Table 2.) Same Gene Same Direction of Differential Expression											
Basal B							Basal A				
Downregulated			Upregulated				Downregulated			Upregulated	
			CCND1	3.3777						CCND1	2.9190
CLDN10	-4.3241						CLDN8	-3.3625			
CYP4X1	-3.7687						CYP4X1	-4.6735			
			E2F7	2.7490						E2F5	2.4467
			E2F8	3.3458							
EIF4EBP1	-2.1679						EIF5A	-2.1794			
HIST1H2BG	-3.7896						HIST1H2BD	-2.1687			
HIST2H2BE	-3.0696										
SUSD2	-4.1535						SUSD2	-2.4498			
			TCEAL8	4.1498						TCEAL9	6.5298

When we compared separately, Basal B- CA vs AA, and Basal A- CA vs AA, we identified the genes given in Table 7. We will refer to these as our candidate genes. The gene, its fold-change, significance value, direction of change (i.e., either high or low in AA) and gene description are noted. The APOE gene is indicated, not because it's a reliable candidate, but to serves as example of a gene that is high in 1 condition and low in another. It's important to point out that both Basal A and Basal B contain differentially expressed TCEAL genes, TCEAL9 in Basal A and TCEAL 8 in Basal B. Both TCEAL genes show lower gene expression levels in African-Americans. The TCEAL genes belong to the same gene family and are functionally related making them candidates as African American associated genes. The functional domain studies discussed later will support this hypothesis. Figure 11 is a graphical representation of Table 7. For our candidate genes. Caucasian expression levels are in gray and African American expression levels are in gold. I want to note that the genes that I selected are candidate genes based on preliminary analysis. The genes will be analyzed further by polymerase chain reaction and western analyses for protein, and if available, tissue microarray to examine protein levels in patient

samples. Based on analyses, there are no associations, pathway signaling or otherwise connecting TCEAL8, TCEAL9, EFDH1 and PEG3.

Table 7: Low and High Differentially Expressed Genes
Notes included in the table indicate if the genes were either low or high in AA cell lines.

Basal-Immunomodulatory/ Subtype A					
	GENE	FOLD CHANGE	p VALUE	AFFY ID	NOTES
	APOE	2.7580	0.042795	203381_s_at	low in AA / Apolipoprotein E / catabolism of triglyceride-rich lipoprotein
	TCEAL9	6.5298	0.0062573	217975_at	low in AA / Transcrip Fac Elongation factor L9 / chr X
Mesenchymal stem-like /Subtype B					
	GENE	FOLD CHANGE	p VALUE	AFFY ID	NOTES
	APOE	-4.5825	0.0023212	203381_s_at	high in AA / Apolipoprotein E / catabolism of triglyceride-rich lipoprotein
	TCEAL8	4.1498	0.020759	224819_at	low in AA / Transcrip Fac Elongation factor L8 / chr X
	EFHD1	-2.4989	0.034537	209343_at	high in AA / EF-hand / Ca++ binding / mitosis
	PEG3	-5.9901	0.043638	209242_at	high in AA / paternally expres / Kruppel C2H2-type zinc finger / p53 apop

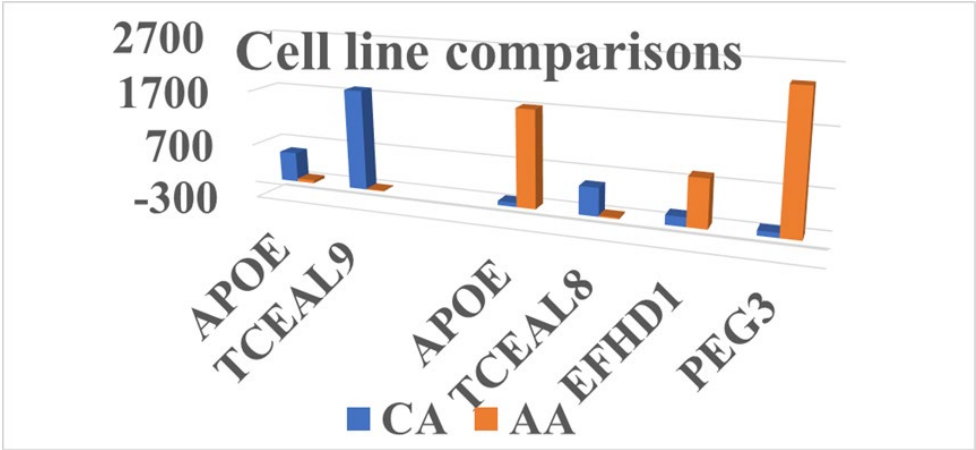


Figure 11: Graph Showing Level of Gene Expression of Our 5 Differentially Expressed Genes

The EFHD1 was also selected as a candidate gene. The gene is an EF-hand Ca++ binding gene associated with mitotic events and synaptic transmission. The gene was chosen mainly because it was identified as differentially expressed in TNBC patient samples; similarly, for PEG3. The patient sample data are presented later. In a separate study, Ruiz-Narváez et al (33) identified what they termed a breast cancer susceptibility locus, 2q37.1 in African American women. The article *Admixture Mapping of African American Women in the AMBER Consortium Identifies New Loci for Breast Cancer and Estrogen-Receptor Subtypes* is the largest of its kind including thousands of patients many which are African American. The authors did not identify genes at this locus. We are interested in this study because the EFHD1 identified in our study is located at 2q37.1. The EFHD1 gene is downregulated in Basal B AA women and might be mutated in these patients leading to loss of heterozygosity. We will continue these analyses.

Cluster Analysis Validating Cell Line Data

We also performed a cluster analysis of our genes (Figure 12). The cell lines included all molecular subtypes of breast cancer. Our candidate genes are designated across the top of the cluster and cell lines are along the side. Based on a particular gene, the program determines and compares the transcript levels of that gene in the various cell lines. All of the cell lines were obtained from the GSE12777 microarray study to eliminate potential bias that might occur when comparing cell lines from different studies. Transcript levels are coded in red or blue. Red represents high transcript levels and blue represents lower levels for a gene. Based on analyses, there are no associations, pathway signaling or otherwise connecting TCEAL8, TCEAL9, EFDH1 and PEG3.

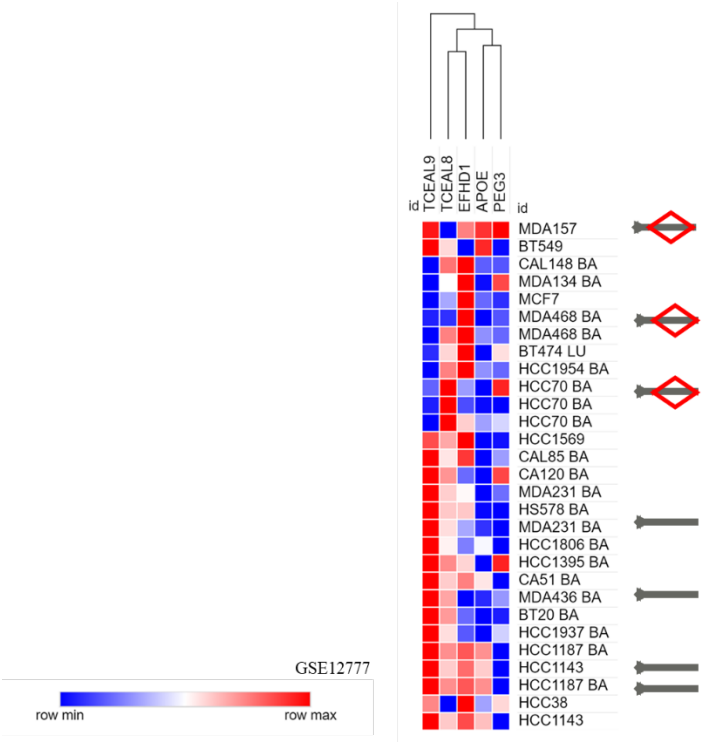


Figure 12: Clustergram of our 5 Differentially Expressed Genes in Diverse Breast Cancer Cell Lines Including Luminal A/B, Her2neu, TNBC, Normal-like

The cell lines and the genes that cluster closer together are more closely related (based on molecular characteristics), those farther apart are more different (31). Interestingly, all of the African American samples clustered in the upper region of our clustergram shown by the top three arrows with red boxes, and all of the Caucasian samples clustered in the lower region of our clustergram shown by the lower four arrows. At first thought, we assumed that Basal A would cluster together, and Basal B would cluster together, irrespective of ethnicity. But the clustergram depends on the genes being analyzed. This particular clustergram is considered a 'supervised analyses' where the investigator determines the genes to process. Based on this analysis, these data validate

that AA cell lines are more closely related to each other. Modifications to the samples or the genes will generate a different response/answer.

T-Test Analysis of Patient Samples

We further validated our cell line data by analyzing the same genes in patient samples. We were able to find a microarray dataset containing AA vs CA TNBC patients. This is not a perfect dataset (for us) because the patients are not characterized by Lehmann classification. They were identified as CA TNBC and AA TNBC. Patient samples were selected from GDS3097 based on ER/Her2 negative receptor status. Nonetheless, we performed the analyses using the usual T-test. Results from this comparison are included in Table 8. The datasets also differ from those described above in that a different microarray platform was used. These authors used the older U133A microarray which contained 22,000 genes and probe-sets compared to the higher density (54,000 probe-set) microarray used for all the previous analyses. The TCEAL genes are not on this array, so they can not be analyzed.

We performed a T-test analysis using the same filter criteria as before (i.e., p-value <0.05- and 2-fold differential gene expression). Two of our candidate genes were identified in this dataset, including the PEG3 and the EFHD1 genes, similar to our cell line data. It's important to reiterate that there are limitations to using this patient sample dataset. The samples are not stratified based on Lehman classification like the cell lines used in our previous analyses. Still PEG3 and EFHD1 are identified as differently expressed in African American patient samples vs Caucasian patient samples. We were encouraged by these findings.

Table 8: Patient Sample T-Test

T-test group analysis of patient samples. Validation of cell line data in actual patient samples (CA red vs AA black)

GSM136394	GSM136400	GSM136412	GSM136387	GSM136391	GSM136406	GSM136413	GSM136416	GSM136417	GSM136418	GSM136420	A-B p-Value	A-B Mean Difference	Feature ID	Gene
94.6	53.4	46.7	80.0	121.0	620.0	470.2	700.6	205.6	260.7	506.7	0.0023815	-2.2434	203962_s	NEBL
64.1	55.4	50.6	4901.0	1594.4	1521.1	557.4	479.5	98.4	2561.1	42.3	0.0037024	-3.5366	205030_at	FABP7
72.1	32.4	32.0	58.1	213.0	441.6	370.6	483.7	192.8	230.0	157.0	0.0043409	-2.4228	209369_at	ANXA3
44.5	39.2	115.6	1797.2	409.5	162.9	1374.0	3267.3	49.3	141.7	1620.0	0.0072156	-3.1714	209242_at	PEG3
57.0	73.4	142.2	416.7	1018.3	1033.8	266.9	376.6	255.5	272.4	106.5	0.0081671	-2.1244	205110_s	FGF13
14.7	23.0	15.4	862.9	121.2	186.7	49.2	68.3	18.7	298.2	14.0	0.011401	-2.4087	205029_s	FABP7
50.7	28.6	19.2	186.9	170.5	319.2	95.9	53.9	752.5	109.9	22.4	0.0138	-2.1342	204913_s	SOX11
43.9	78.7	122.5	56.9	154.7	866.2	489.5	324.4	351.3	321.6	583.6	0.01462	-2.0340	212328_at	LIMCH1
779.7	2611.5	1035.8	87.6	213.8	187.8	194.3	180.3	166.6	763.2	444.0	0.017942	2.5006	209396_s	CHI3L1
165.7	493.1	284.6	740.0	819.4	1551.9	735.7	1459.1	2677.6	1904.7	652.3	0.02331	-2.0274	209343_at	EFHD1
83.6	27.1	36.8	94.0	467.0	376.1	112.6	182.6	269.9	125.2	147.6	0.024012	-2.1195	204875_s	GMDS

.

Further Analyses of the TCEAL Family of Genes:

The TCEAL genes belong to a family of transcription elongation factor genes. The gene family is characterized as having a brain expressed linked family (BEX) functional domain. The BEX1 domain is thought be associated with genes involved in cell cycle and signaling, neuronal differentiation and general responses to external signaling events. All TCEAL genes contain the BEX domain. The TCEAL genes are small with slight variations in sequence and length and interestingly they are located on the chromosome X. Our data show TCEAL8 and 9 downregulated in AA TNBC. Interestingly, a study by Huo et al. identified TCEAL7 associated with TNBC. The study "*Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas* was performed by Huo and Charles Perou (34). The authors identified 142 genes as differentially associated with AA women and TCEAL7 was one of the genes. TCEAL7 is also a member of the TCEAL gene family.

STRING Analyses of TCEAL Proteins

We identified both TCEAL genes downregulated in AA TNBC cell lines. TCEAL8 was downregulated in Basal B, and TCEAL9 was downregulated in Basal A cell lines. We suggest that TCEAL8 might function in Basal B, and TCEAL9 performs the same function in Basal A. To examine if the proteins have been characterized as functionally similar, we performed the STRING assay (32) string.db.org) (8) . STRING program interrogates millions of data points based on (a) if particular genes have been sited in the same publication and (b) if the genes have been experimentally determined to interact via any experimental method. The STRING analyses demonstrate a close relationship between TCEAL8 and TCEAL9 (WBP5). The WBP5 gene is another designation for TCEAL9. Analysis determined that TCEAL 8 and TCEAL 9 genes from our list of candidate genes are functionally related. It could be that TCEAL8 involved in signaling mechanisms in Basal B and TCEAL9 is involved in signaling mechanisms in Basal A. STRING displays a low (i.e., high confidence, statistically significant value) for TCEAL8 and TCEAL9 involvement in downregulation of elongation functions. These results must be analyzed further, and experimental comparisons performed via polymerase chain reaction. TCEAL7 was not found associated with TCEAL8 and 9. There are no associations, signaling or otherwise connecting TCEAL8, TCEAL9, EFDH1 and PEG3.

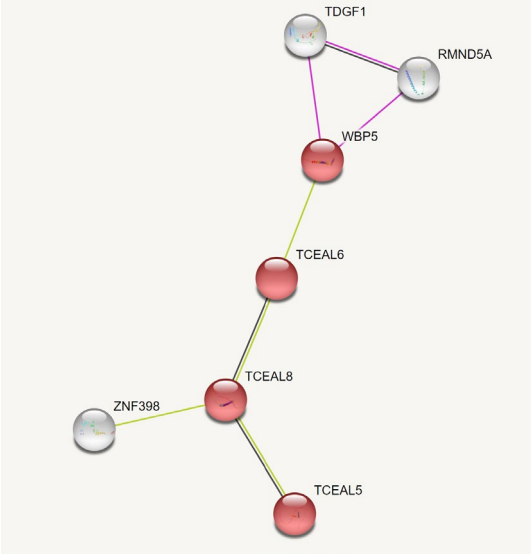


Figure 13: **STRING to Demonstrate Functional Relationship between TCEAL9 (WBP5) and TCEAL8**

Red balls represent genes involved in decreased expression of transcription elongation. Significant strength.

publication	(year) title	count in network	strength	false discovery rate
PMID:23372750	(2013) Decreased expression of transcription elongation fa...	4 of 14	2.9	5.84e-05

TCEAL Comparison Analysis Using BLAST, Conserved Domain and Multialin BLAST Analysis of TCEALs

The Multalin program is used to perform sequence alignment. TCEAL 8 and TCEAL 9 genes were subjected to analyses at the following website (http://multalin.toulouse.inra.fr/multalin/) (28) Figure 14. The multalin software allows simultaneous alignment of several biological sequences so that similarities in the sequences are juxtaposed (28). These comparisons show NP_699164.1 (TCEAL8) and NP_057387

(TCEAL9) are closely aligned in the BEX family domain (at ~position 90). The TCEAL7 protein differs at this domain by 2 amino acids (instead of RV there are amino acids GL).

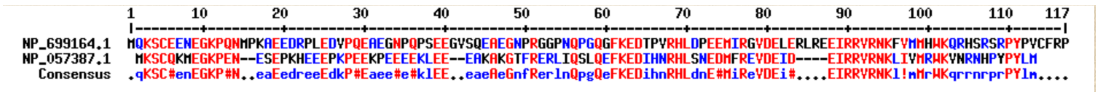


Figure 14: Multalin data Comparing the Sequence Alignment of TCEAL 8 and 9 Protein
The common functional domain is amino acids positions 88-96.

As mentioned, from a functional standpoint TCEAL8 and TCEAL9 genes are closely related, they are both on the X-chromosome, both elongation factors, and when you compare their functional domains, they are the same. The functional domain of a protein will define how proteins work, which might explain why both TCEAL 8 and TCEAL9 show low expression levels in our cell line data. We performed a BLAST analysis of TCEAL8 to determine genes with similar homology (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (27). With TCEAL8 as the input gene, the program automatically searches for sequence homology based oh homo sapiens (because we limited the analyses to humans). Table 9 show the TCEAL genes are the highest 'confidence' genes similar to TCEAL8 which include TCEAL9, TCEAL7 and two TCEAL1 genes. We recognize that for any particular protein a number of functional domains exist. So for TCEALs, there are likely uncharacterized domains present. Maybe its coincidental that the TCEALs are downregulated in the AA cancers and the function of the gene is independent of functions related to BEX. These data only serve as preliminary

data, and we understand that additional data analyses and definitely experimental studies must be done.

Table 9: Conserved Domain Data of TCEALS
Conserved domain data showing that TCEAL 7 and 9 have a high association to TCEAL 8

Description	Query Cover	E Value	Ident	Accession
transcription elongation factor A protein-like 8 (Homo sapiens]	100%	2.00E-81	100%	NP 699164.1
transcription elongation factor A protein-like7 [Homo sapiens]	96%	3.00E-21	46%	NP 689491 1
transcription elongation factor A protein-like 9 [Homo sapiens]	94%	7.00E-13	41%	NP 057387.1
ReName: Full=Transcription elongation factor A protein-like 1-Short=TCEA-like protein 1; AltName:*Full=Nuclear phosphoprotein p21/SIIR: AltName: Full=Tr 53.9 53.9	45%	3.00E-09	53%	015170 2
Transcription elongation factor A protein-like 1r Homosapiensl	45%	3.00E-09	53%	NP 0047712
pp21 (Homo sapiens]	45%	5.00E-07	49%	AAA60149,1
quanine nucleotide binding_protein (G protein), beta polypeptide 2-like 1 isofor CRA h (Homo sapiens]	37%	0.01	36%	EAW53699.1

.

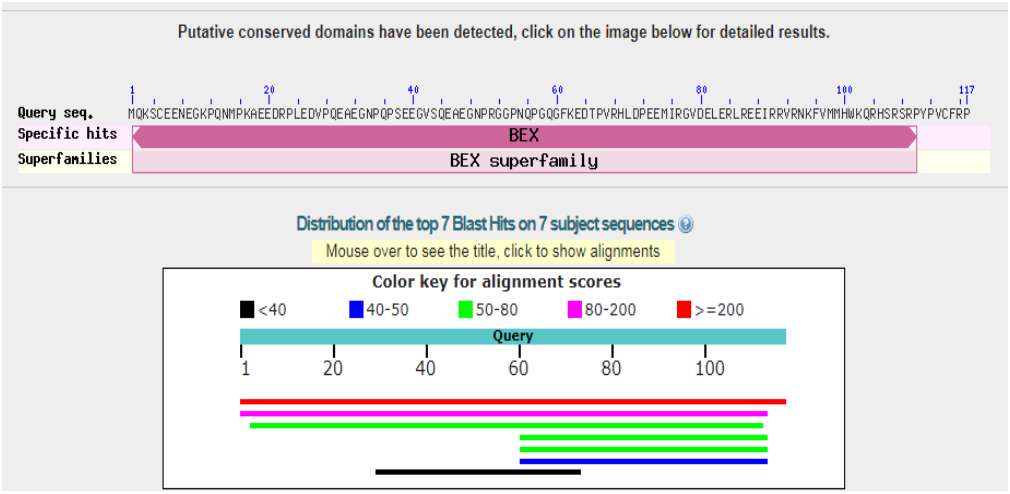


Figure 15: Domain Analyses of TCEAL Family Genes
The top 5 lines include sequence similarity between TCEAL8, TCEAL7, TCEAL9, 2 similar-TCEAL1.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

We have identified 4 different genes (excluding APOE) that are differentially expressed meaning they show different expression levels in African American women cell lines to Caucasian TNBC cell lines. We believe that these differences in expression levels are significant. We validated our genes using a number of different methods, so we have confidence in our discovery. At this point the study is exclusively a bioinformatic study, but we have plans to expand the study and include experimental transcript and protein expression studies as validation. Whether these data will validate experimentally remains to be determined.

.

REFERENCES

1. What Is Cancer? (2021, May 5). National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/what-is-cancer
2. Sarkar, S., Horn, G., Moulton, K., Oza, A., Byler, S., Kokolus, S., & Longacre, M. (2013). Cancer Development, Progression, and Therapy: An Epigenetic Overview. International Journal of Molecular Sciences, 14(10), 21087–21113. https://doi.org/10.3390/ijms141021087
3. National Institutes of Health (US); Biological Sciences Curriculum Study. NIH Curriculum Supplement Series [Internet]. Bethesda (MD): National Institutes of Health (US); 2007. Understanding Cancer.
4. Cancer. (2022, February 3). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cancer
5. Pietrangelo, A. (2022, January 6). Early Signs and Symptoms of Breast Cancer. Healthline. https://www.healthline.com/health/breast-cancer/warning-signs
6. Basumallik N, Agarwal M. Small Cell Lung Cancer. [Updated 2021 Jul 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-.
7. Recio-Boiles A, Cagir B. Colon Cancer. [Updated 2022 Jan 24]. In: StatPearls [Internet]. Treasure Island (FL): Stat Pearls Publishing; 2022 Jan-.

8. Leslie SW, Soon-Sutton TL, Sajjad H, et al. Prostate Cancer. [Updated 2022 Feb 14]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-.

9. Sharma, G., & Dave, R. (2010). VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW. J Adv Pharm Technol Res., 1(2), 109–126.

10. Alkabban FM, Ferguson T. Breast Cancer. [Updated 2021 Aug 7]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-.

11. NCI Dictionary of Cancer Terms. (2022). National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/breast-duct

12. Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk Factors and Preventions of Breast Cancer. International Journal of Biological Sciences, 13(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

13. Akram, M., Iqbal, M., Daniyal, M., & Khan, A. U. (2017). Awareness and current knowledge of breast cancer. Biological Research, 50(1). https://doi.org/10.1186/s40659-017-0140-9

14. How Is Breast Cancer Treated? (2020, September 14). Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/treatment.htm

15. Nounou, M. I., ElAmrawy, F., Ahmed, N., Abdelraouf, K., Goda, S., & Syed-Sha-Qhattal, H. (2015). Breast Cancer: Conventional Diagnosis and Treatment

Modalities and Recent Patents and Technologies. Breast Cancer: Basic and Clinical Research, 9s2, BCBCR.S29420. https://doi.org/10.4137/bcbcr.s29420

16. Lehmann. B.D., Pietenpol JA. Clinical implications of molecular heterogeneity in triple negative breast cancer. *Breast (Edinburgh, Scotland)*. 24: S36-40. PMID 26253813 DOI: 10.1016/J.Breast.2015.07.009

17. Types of Breast Cancer. (2021, February 3). WebMD. https://www.webmd.com/breast-cancer/breast-cancer-types

18. Weigelt, B., F.C. Geyer, and J.S. Reis-Filho, Histological types of breast cancer: how special are they? Mol Oncol, 2010. 4(3): p. 192-208.

19. Yersal, O. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. World Journal of Clinical Oncology, 5(3), 412. https://doi.org/10.5306/wjco.v5.i3.412

20. Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. Journal of Clinical Investigation, 121(7), 2750–2767. https://doi.org/10.1172/jci45014

21. The American Cancer Society, “About Breast Cancer”. *Cancer.org.* January 25, 2022

22. Siddharth, S., & Sharma, D. (2018). Racial Disparity and Triple-Negative Breast Cancer in African American Women: A Multifaceted Affair between Obesity, Biology, and Socioeconomic Determinants. Cancers, 10(12), 514. https://doi.org/10.3390/cancers10120514

23. Ring BZ, Hout DR, Morris SW, Lawrence K, Schweitzer BL, Bailey DB, **Lehmann BD**, Pietenpol JA, Seitz RS. Generation of an algorithm based on minimal gene sets to clinically subtype triple negative breast cancer patients. *Bmc Cancer*. 16: 143. PMID 26908167 DOI: 10.1186/S12885-016-2198-0

24. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies **Brian D Lehmann** 1, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, Jennifer A Pietenpol

25. Brian D. Lehmann, Ph.D. | Pietenpol & Lehmann Lab. (2022). Vanderbilt University Medical Center. https://www.vumc.org/pietenpol-lab/person/brian-d-lehmann-phd

26. microarray | Learn Science at Scitable. (2014). Nature.Com. https://www.nature.com/scitable/definition/microarray-202/?error=cookies_not_supported&code=124e9d1a-1b31-4384-bcab-853d1bc4f941

27. Jovanović B, Sheng Q, Seitz RS, Lawrence KD, Morris SW, Thomas LR, Hout DR, Schweitzer BL, Guo Y, Pietenpol JA, **Lehmann BD**. Comparison of triple-negative breast cancer molecular subtyping using RNA from matched fresh-frozen versus formalin-fixed paraffin-embedded tissue. *Bmc Cancer*. 17: 241. PMID 28376728 DOI: 10.1186/S12885-017-3237-1

28. Lee YC, Chen YJ, Wu CC, Lo S, Hou MF, Yuan SS.Gynecol Oncol. 2012 Jun;125(3):742-50. doi: 10.1016/j.ygyno.2012.02.032. Epub 2012 Feb 24.

29. Field LA, Love B, Deyarmin B, Hooke JA, Shriver CD, Ellsworth RE. Cancer. 2012 Mar 1;118(5):1334-44. doi: 10.1002/cncr.26405. Epub 2011 Jul 28.

30. Sturtz LA, Melley J, Mamula K, Shriver CD, Ellsworth RE.BMC Cancer. 2014 Feb 4; 14:62. doi: 10.1186/1471-2407-14-62.

31. Lund MJ, Trivers KF, Porter PL, Coates RJ, Leyland-Jones B, Brawley OW, Flagg EW, O'Regan RM, Gabram SG, Eley JW.Breast Cancer Res Treat. 2009 Jan;113(2):357-70. doi: 10.1007/s10549-008-9926-3. Epub 2008 Mar 7.

32. *Cluster Analysis: Definition and Methods*. (2022, February 12). Qualtrics. https://www.qualtrics.com/experience-management/research/cluster-analysis/

33. *STRING: functional protein association networks*. (2022). STRING. https://string-db.org/

34. Ruiz-Narváez, E. A., Sucheston-Campbell, L., Bensen, J. T., Yao, S., Haddad, S., Haiman, C. A., Bandera, E. V., John, E. M., Bernstein, L., Hu, J. J., Ziegler, R. G., Deming, S. L., Olshan, A. F., Ambrosone, C. B., Palmer, J. R., & Lunetta, K. L. (2016). Admixture Mapping of African American Women in the AMBER Consortium Identifies New Loci for Breast Cancer and Estrogen-Receptor Subtypes. Frontiers in Genetics, 7. https://doi.org/10.3389/fgene.2016.00170

35. Huo, D., Hu, H., Rhie, S. K., Gamazon, E. R., Cherniack, A. D., Liu, J., Yoshimatsu, T. F., Pitt, J. J., Hoadley, K. A., Troester, M., Ru, Y., Lichtenberg, T., Sturtz, L. A., Shelley, C. S., Benz, C. C., Mills, G. B., Laird, P. W., Shriver, C. D., Perou, C. M., & Olopade, O. I. (2017). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncology, 3(12), 1654. https://doi.org/10.1001/jamaoncol.2017.0595