CHARACTERIZING BASAL-LIKE TRIPLE NEGATIVE BREAST CANCER USING GENE EXPRESSION ANALYSIS: A DATA MINING APPROACH

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Biomedical Engineering

By

QAMAR ALSABI

B.S.B.E., Wright State University, 2017

2019
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

November 22, 2019

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY QAMAR ALSABI ENTITLED CHARACTERIZING BASAL-LIKE TRIPLE NEGATIVE BREAST CANCER USING GENE EXPRESSION ANALYSIS: A DATA MINING APPROACH BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Biomedical Engineering.

____

.Jaime E Ramirez-Vick, Ph.D.
Thesis Director

____

John C. Gallagher, Ph.D.
Chair, Biomedical, Industrial, and
Human Factor Engineering.

Committee on Final Examination:

____

Nasim Nosoudi, Ph.D.

____

Amir Zadeh, Ph.D.

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Alsabi Qamar. M.S.B.M.E., Department of Biomedical, Industrial, and Human Factor Engineering, Wright State University, 2019. Characterizing Basal-Like Triple Negative Breast Cancer using Gene Expression Analysis: A Data Mining Approach.

Triple-negative breast cancer (TNBC) is characterized by the absence of expression of the estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 (HER2). Therefore, TNBC is unresponsive to targeted hormonal therapies, which limits treatment options to nonselective chemotherapeutic agents. Basal-like breast cancers (BLBCs) represent a subset of about 70% of TNBCs, more frequently affecting younger patients, being more prevalent in African-American women and significantly more aggressive than tumors of other molecular subtypes, with high rates of proliferation and extremely poor clinical outcomes. Proper classification of BLBCs using current pathological tools has been a major challenge. Although TNBCs have many BLBC characteristics, the relationship between clinically defined TNBC and the gene expression profile of BLBC is not fully examined. The purpose of this study is to assemble publicly-available TNBC gene expression datasets generated by Affymetrix gene chips and define a set of genes, or gene signature, that can classify TNBC samples between BLBC and Non-BLBC subtypes. We compiled over 3,500 breast cancer gene expression profiles from several individual publicly available datasets and extracted Affymetrix gene expression data for 580 TNBC cases. Several popular data mining methods along with dimensionality reduction and feature selection techniques were applied to the resultant dataset to build predictive models to understand molecular characteristics and mechanisms associated with BLBCs and to classify them more accurately according to important features extracted through microarray data analysis of BLBC and Non-BLBC cases. Our result can lead to proper identification and diagnosis of BLBCs, which can potentially direct clinical implications by dictating the most effective therapy.

Table of Contents

TABLE OF FIGURES

LIST OF TABLES

Introduction

Triple-negative breast cancer (TNBC) constitutes approximately 20%-25% of all breast cancer cases with poor prognosis.[1] TNBC is defined as the lack of specific breast-cancer-associated receptors, mainly progesterone (PR), estrogen (ER), and human epidermal growth factor (HER2). As a result, due to the lack of targets TNBC is unresponsive to targeted hormonal therapies, which limits treatment options to nonselective chemotherapeutic agents.[2]

Recent technological advances allow for high throughput profiling of biological systems at the molecular level in a cost-efficient manner. The relatively low cost of data generation is leading us to the "Big Data Era". Today big data can be created out of small data and the combination of datasets from various sources is a major aspect of "big data". The availability of such large datasets provides unprecedented opportunities for data mining, deep learning, and integrative analysis over various layers of data which set the goal to link all the molecular information and translate it back into meaningful information in precision medicine, systems biology, molecular physiology or pathophysiology.

Translational modeling is not new to cancer research. Predictive modeling has been applied in clinical domains and into a wide variety of problems in breast cancer such as

[1] KR Bauer. Descriptive analysis of estrogen receptor (ER)- negative, progesterone receptor (PR)-negative. And HER2-negative invasive breast cancer. The so-called triple-negative phonotype. (population-based study from the California cancer Registry). 1721-1728

[2] Ibid

diagnosis[3],survivability[4], prognosis[5], susceptibility[6] and recurrence[7]. However, the extent to which microarray data can improve the diagnosis of BLBC cancer has not been fully examined.

The purpose of this analysis is to assemble publicly-available TNBC gene expression datasets generated on Affymetrix gene chips and define a set of genes, or gene signature, that can classify TNBC between the basal-like breast cancer (BLBC) and Non-basal-like breast cancer (Non-BLBC) subtypes. A proper diagnosis of BLBC will have clinical implications by dictating the most effective therapy.

The approach that used to characterize basal-like triple negative breast cancer is data mining approach using supervised analysis (i.e., classification). Eight data mining techniques were used to classify basal-like triple negative breast cancer include Neural Network, Decision tree, Logistic Regression, Support Vector Machine, Least Angle Regression, Gradient, Random Forrest, and Bayesian Classifier.

[3] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*(2), 3240-3247

[4] D Delen., G Walker., & A Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine, 34*(2), 113-127 (2005).

[5] Chen, A. H., & Yang, C. (2012). The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data. *Expert Systems with Applications, 39*(5), 4785-4795.

[6] Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn Jr, C. E., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer, 116*(14), 3310-3321.

[7] Kim, W., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., . . . Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer, 15*(2), 230-238.

Background

Breast Cancer (BC)

Environmental and genetic factors are the main causes of Breast Cancer (BC), due to the accumulation of mutations in essential genes.[8] In developed countries, BC is the most common cancer in women, being the cause of death in approximately 20% of females diagnosed.[9] In the case of African-American women under the age of 50 years of age 39% of the diagnosed BC cases are of the TNBC type, while they only represent 16% in Caucasian women.[10] Based on global gene expression analyses, four molecular subtypes of BC have been identified, mainly, luminal A, luminal B, HER2-enriched and basal-like. These subtypes have shown to be significantly different in terms of their baseline prognosis, age at diagnosis, risk factors and response to therapies. Among these types, basal-like breast cancer is of great interest to investigators and clinicians due to its poor prognosis, high frequency, limited targeted therapies.[11]

Triple Negative Breast Cancer (TNBC)

TNBC is defined as a type of BC which shows the absence of the three common BC biomarkers, PR, ER, and HER2.[12] TNBC tends to be more aggressive compared to other BC types. In addition, the chance of early recurrence is high, due to the absence of the ER.[13]

[8] Nathanson K.N, Wooster R, Weber B.L. Breast cancer genetics: What we know and what we need. 552-556

[9] F Macdonald, Ford CHJ, AG Casson. Breast cancer. In 'Molecular Biology of Cancer'.139-63

[10] LA Carey, CM Perou and CA Livasy. Race, breast cancer subtypes, and survival in the Carolina breast cancer study. 2492-502

[11] Prat Aleix, A Barbara, C Maggie, A Carey, C Lisa and P Charles. Molecular Characterization of Basal-Like and None-Basal-Like Triple-negative Breast Cancer. 123-133

[12] KR Bauer. (2007). 1721-1728

[13] Ibid

The absence of the BC-specific targets ER, PR, and HER2, limits the treatment options for TNBC. These include hormone therapies, anti-HER2 targeted therapies, endocrine (tamoxifen, aromatase inhibitor inhibitors) therapy, and trastuzumab (anti-HER2). TNBC cases only achieve 19% clinical-complete-response to chemotherapy.[14] This leaves as the only treatment option available for TNBC, cytotoxic chemotherapy.[15]

Although TNBC has many BLBC characteristics, the relationship based on the gene expression is not completely clear, where not all TNBC cases fall into the BLBC subtype.[16]

Basal-like Breast Cancer (BLBC)

BLBC represents approximately 15-20% of breast cancer cases,[17] and is defined as being ER negative, PR negative, cytokeratin 5/6 positive and/or HER2 positive.[18] It mainly occurs at an early age, showing an aggressive clinical outcome, presence of distant metastases, especially within the first five years after the diagnosis, showing poor prognosis, and a high mortality rate.

BLBC subtype of TNBC

Based on the protein profile, 53-84% of TNBC cases are diagnosed as BLBC.[19] Another study reported that 6 of 31 (19.4%) triple- negative breast tumors were classified as Non-

[14] J Choi, WH Jung and JS Koo. Clinicopathologic features of molecular subtypes of triple negative breast cancer based immohistochemical markers. 1481-93

[15] C liedtke, C Mazouni, KR Hess, F Tordai, JA Mejia, WF symmans, AM Gonzalez-Angulo, B Hennessy and M Green. Response to neoadjuvant therepy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol.* 1275-1281

[16] P Boyle. Triple-negative breast cancer: epidemiological considerations

[17] Badowsha-Kozakiewicz and Budzik: Immunohistochemical charactristics of basal-like breast cancer.436-443

[18] Ibid

[19] Y Liu, T Xin and QY Jiang. CD147, MMP9 expression and clinical significance of basal-like breast cancer.

BLBC, while 15 out of 207 (6.3%) non-triple-negative tumors showed basal cytokeratin biomarkers.[20] A previous investigation showed that 69.7% of TNBC were classified as BLBC.[21] Until now, there is no accepted definition to classify BLBC. To improve the criteria for defining BLBC, some studies included microarray-based expression profiling data, and panels of immunohistochemical surrogates, which yielded a definition that included cancer tissue (1) with the absence of ER, PR, and HER2 expression (i.e., triple-negative); (2) expressing one or more high-molecular-weight/basal cytokeratin (i.e., CK5/6, CK14, or CK17), which are usually expressed in the basal epithelial layer of skin and airways, but are also expressed in some breast carcinomas; (3) absence of ER and HER2 expression in conjunction with CK5/6 and or epidermal growth factor EGFR; (4) absence of ER, PR, and HER2 expression in conjunction with CK5/6 and/or EGFR.[22]

Diagnostic Difficulties

Unlike other subtypes of BC, the BLBC subtype seems not to correlate with the size of the primary tumor and the presence of regional lymph node metastases.[23] However, there are a variety of immunohistochemical markers that can be used to identify BLBC, such as cytokeratins (CK5/6, CK14 or CK17), EGFR, smooth muscle actin (SMA), p63, p-cadherin, ki-67, p53 or c-kit antigen with concomitant lack of ER, PR, HER2 and “luminal” cytokertins (CK8, CK18, CK19) expression.[24] BLBC shows higher genome instability compared to other BC subtypes. Therefore, there is no particular set of markers that

[20] DS Tan, C Marchio, RL Jones. Triple negative breast cancer: molecular profiling and progistic impact in adjuvant anthracycline-treated patients. 27-44

[21] Rody.(2011). A clinically relevant gene signature in triple negative and basal-like breast cancer.

[22] MC Cheang. Basal-like breast cancer defined by five biomarkers has superior value than triple-negative phenotype. 1368-1376

[23] Badowsha-Kozakiewicz and Budzik (2016).436-443

[24] Ibid

explicitly define BLBC.[25] However, a more detailed classification of TNBC tumors needs to be established because of the variability shown within this type based on molecular studies[26]. Moreover, to define better prognostic biomarkers and therapeutic alternatives, further investigations are needed to better classify TNBC, BLBC, and Non-BLBC tumors.

Biomarkers in TNBC and BLBC

A biomarker is a biological molecule that serves as a sign for normal biological processes or conditions or signals the presence of an abnormal process, condition, and thus, the presence of a biological defect, risk to a particular ailment, or an actual disease. Researchers have explored biomarkers for selected types of cancer to aid in prevention or risk assessment, diagnostic, and treatment or management[27]. Nonetheless, the existing body of literature remains unorganized when it comes to biomarkers for more specific types of cancer, such as in the case of TNBC as BLBC, TNBC as Non-BLBC, or Non-TNBC as BLBC.

Two of the earliest identified biomarkers for general breast cancer are BRCA1 and BRCA2, which are related tumor suppressor genes responsible for repairing DNA or destroying cells if DNA damage is irreparable. Damage in either of these two genes, due to specific heritable mutations, increases the risk of cancer in breast tissue, as well as in ovarian and blood tissue due to the loss DNA repair capacity[28] Although, both BRCA1 and BRCA2 are important biomarkers for susceptibility to breast and other types of cancer,

[25] Badowsha-Kozakiewicz and Budzik.(2016).436-443.

[26] Ibid

[27] Verma, Mukesh, and Upender Manne. "Genetic and Epigenetic Biomarkers in Cancer Diagnosis and Identifying High Risk Populations." *Critical Reviews in Oncology/Hematology* 60, no. 1 (October 2006): 9–18.

[28] Friedenson, Bernard. "The BRCA1/2 Pathway Prevents Hematologic Cancers in Addition to Breast and Ovarian Cancers." *BMC Cancer* 7, no. 1 (August 6, 2007). doi:10.1186/1471-2407-7-152.

their capacity in defining susceptibility and presence of TNBC and BLBC, has been very limited.

Several review studies have identified other alternative biomarkers. For instance, one study[29] did a comprehensive review of PubMed and conference databases to evaluate the literature concerning TNBC biomarkers. The study listed the following biomarkers: epidermal growth factor receptor (EGFR), vascular endothelial growth factor, c-Myc, C-kit, basal cytokeratins, poly(ADP-ribose) polymerase-1, p53, tyrosinase kinases, m-TOR, heat and shock proteins, and TOP-2A. The same study[30] noted that the absence of estrogen receptors or ER, progesterone receptors or PR, and HER-2/neu receptors are distinctive biomarkers for BLBC and they represent 80% of TNBC cases. Other studies[31] have identified additional biomarkers for BLBC, including EGFR and cytokeratin CK 5/6, which are keratin proteins that serve as essential components of intermediate filaments that help cells resist mechanical stress.

It is also important to note the differential expression of other keratin proteins is also seen in both TNBC and BLBC. For instance, the differential expression of CK7, CK8, CK18, and CK19 was observed in more than 90 percent of all breast carcinomas. In addition, the expression of CK5/6, CK14, and CK20 positively correlated with a high tumor grade.[32] Another study[33] that analyzed 11 TNBC tumors, identified the specific occurrence of

[29] Yadav, Budhi S. "Biomarkers in Triple Negative Breast Cancer: A Review." *World Journal of Clinical Oncology* 6, no. 6 (2015): 252. https://doi.org/10.5306/wjco.v6.i6.252.

[30] Ibid

[31] Cheang, M. C.U., D. Voduc, C. Bajdik and S. Leung, 1368–76.

[32] Shao, M.-M., Chan, S. K., Yu, and A. M. C. (2012). Keratin expression in breast cancers. *Virchows Archiv*, 461(3), 313–322.

[33] Kuroda, Naoto, Masahiko Ohara, Kaori Inoue and Keiko Mizuno. "The Majority of Triple-Negative Breast Cancer May Correspond to Basal-like Carcinoma, but Triple-Negative Breast Cancer Is Not Identical to Basal-like Carcinoma." *Medical Molecular Morphology* 42, no. 2 (June 2009): 128–31.

keratins. For example, eight of the tumors were positive for basal markers, CK5 and CK17, six of which also were also positive for CK14. The study concluded that the use of combination immunohistochemistry, which included CK5, CK14, and CK17, could contribute to the detection of basal-like carcinoma.

Overexpression of the protein-coding gene ID4, which is associated with the regulation of many cellular processes during both prenatal development and tumorigenesis, and TP53, which prevents cancer formation through tumor suppression and genome mutation prevention, have been linked to BLBC. The high expression of Ki67 mRNA has also been associated with the high proliferation of BLBC subtype.[34] Note that this nuclear protein has been suggested to play a necessary role in cellular proliferation, as well as in the ribosomal RNA transcription.

Other biomarkers have been identified to indicate both prognosis and therapeutic response to TNBC and BLBC. For example, secreted frizzled related protein 1 or SFRP1 has been found to be a potential molecular marker for response to chemotherapy and potential prognostic marker[35] and an increased secretion of this protein has been associated with higher expression in basal-like cancer cell lines. Thus, one study concluded that SFRP1 is correlated with both an aggressive form of breast cancer and positive response to neoadjuvant chemotherapy.[36]

[34] Yadav, 252-263

[35] Huelsewig, Carolin, Christof Bernemann and Christian Ruckert. "Abstract 920: Secreted Frizzled Related Protein 1 (SFRP1) as Potential Regulator of Chemotherapy Response for Patients with Triple Negative Breast Cancer (TNBC)." In Clinical Research (Excluding Clinical Trials). *American Association for Cancer Research*, 2014.

[36] Bernemann, Christof, Carolin Hülsewig and Christian Ruckert et al. "Influence of Secreted Frizzled Receptor Protein 1 (SFRP1) on Neoadjuvant Chemotherapy in Triple Negative Breast Cancer Does Not Rely on WNT Signaling." *Molecular Cancer* 13, no. 1 (2014): 174.

Another notable example centers on the interplay between two genes that are both indicative of cancer development and prognosis. For instance, one study[37] revealed that microRNA-26 appears to inhibit the metastasis of TNBC, by targeting transmembrane 4 L6 family member 1 or TM4SF1. Note that TM4SF1 expression in breast cancer tissues is higher than that in adjacent normal breast tissues. Furthermore, the expression level of TM4SF1 in MDA-MB-231 cells was associated with the metastatic tendency of TNBC. Nonetheless, the overexpression of miR-206 in the same MDA-MB-231 cells appears to down-regulate TM4SF1.

The interplay with forkhead box C1 or FOXC1 and chemokine receptor-4 or CXRC4, also affect TNBC and BLBC prognosis and metastasis. Specifically, FOXC1 overexpression boosts TNBC metastasis by activating the transcription of CXRC4. However, in a zebrafish tumor model, either AMD3100 or siRNA in MDA-MB-231 cells can inhibit CXRC4 by under-expressing FOXC1.[38]

Moreover, primary breast cancer tissues and its derived cell lines and, particularly, in TNBC tissues and cell lines have up-regulated microRNA-761. The overexpression of exogenous microRNA-761 augmented the TNBC cell proliferation, colony formation, migration, and invasion in vivo. Essentially, microRNA-761 represses the expression of TRIM29, thus inducing aggressive phenotypes in TNBC cells. On the other hand, the overexpression of TRIM29 reversed the proliferative and invasive capacities of TNBC

[37] Fan, Chunni, Ning Liu, Dan Zheng, Jianshi Du, and Keren Wang. “MicroRNA-206 Inhibits Metastasis of Triple-Negative Breast Cancer by Targeting Transmembrane 4 L6 Family Member 1.” *Cancer Management and Research Volume* 11 (July 2019): 6755–64.

[38] Pan, Hongchao, Zhilan Peng, Jiediao Lin and Xiaosha Ren. “Forkhead Box C1 Boosts Triple-Negative Breast Cancer Metastasis through Activating the Transcription of Chemokine Receptor-4.” *Cancer Science* 109, no. 12 (November 18, 2018): 3794–3804.

cells[39]. Note that microRNA-761 is a non-coding RNA that affects the translation and stability of mRNAs. TRIM29 or tripartite motif-containing protein 2 encodes a gene belonging to the TRIM protein family and may act as a regulatory factor involved in carcinogenesis and/or differentiation. However, a high level of another TRIM protein known as TRIM28 with TNBC. The down-regulation and depletion of this protein reduced the ability of TNBC cells to induce tumor growth when injected subcutaneously, thereby resulting in a significant reduction of tumor growth.[40]

Another gene linked to the proliferation of cancerous mammary cells is actin-related protein 2/3 complex or ARPC2. One study[41] screened the Oncomine database and found micro-profiling studies that linked the overexpression of ARPC2 proteins to cancerous cell lines. Furthermore, they found a unique link between ARCP2 overexpression and invasion, apoptosis, and proliferation of mammary carcinoma cells, including tumor size, lymph node metastasis, tumor grade, poor prognosis and response to treatment. Another study[42] showed that that the up-regulation of stearoyl-CoA desaturase 1 or SCD1 was associated with shorter survival in breast cancer patients. A study[43] of specific TNBC subtypes, noted that SCD1 inhibition had been reported to reduce the proliferation and survival of cancer cells, thereby suggesting a new targeted therapeutic approach.

[39] Guo, Guang-Cheng, Jia-Xiang Wang, Ming-Li Han, Lian-Ping Zhang, and Lin Li. "microRNA-761 Induces Aggressive Phenotypes in Triple-Negative Breast Cancer Cells by Repressing TRIM29 Expression." *Cellular Oncology* 40, no. 2 (January 4, 2017): 157–66.

[40] Czerwińska, Patrycja, Parantu K. Shah and Katarzyna Tomczak et al. "TRIM28 Multi-Domain Protein Regulates Cancer Stem Cell Population in Breast Tumor Development." *Oncotarget* 8, no. 1, 10 Nov 2016.

[41] Cheng, Zhongle, Wei Wei, Zhengshen Wu, Jing Wang, Xiaojuan Ding, Youjing Sheng, Yinli Han, and Qiang Wu. "ARPC2 Promotes Breast Cancer Proliferation and Metastasis." *Oncology Reports*, April 12, 2019.

[42] Holder, Ashley M., Ana M. Gonzalez-Angulo and Huiqin Chen. "High Stearoyl-CoA Desaturase 1 Expression Is Associated with Shorter Survival in Breast Cancer Patients." *Breast Cancer Research and Treatment* 137, no. 1 (December 4, 2012): 319–27.

[43] Hosokawa, Yuko, Noritaka Masaki and Shiro Takei. "Recurrent Triple-Negative Breast Cancer (TNBC) Tissues Contain a Higher Amount of Phosphatidylcholine (32:1) than Non-Recurrent TNBC Tissues.", no. 8 (August 23, 2017).

The mRNA expressions of several S100 family of genes have been associated with malignancies in human breast tissue. In the case of TNBC, an analysis[44] using the Kaplan-Meier plotter database revealed that S100P expression is significantly associated with poor survival in TNBC patients. The abundance of mRNA S100P is indicative of poor overall survival of these patients. Another study[45] involved silencing the pi subunit of the GABA(A) receptor or GABRP in vitro. Results revealed a decreased GABRP tumorigenic potential and migration to be concurrent with alterations in the cytoskeleton of basal-like cell lines, by reducing cellular protrusions and expression of several cytokeratin proteins related with BLBC, such as KRT5, KRT6B, KRT14, and KRT17.

The identification of genetic biomarkers for TNBC and BLBC should involve considering the following three points. First, focus on genes that can determine the existence and early-stage development of TNBC and BLBC. Examples of these genes include the under-expressed BRCA1 and BRCA2 tumor suppressor genes, as well as the under-expressed ER, PR, and HER2 receptors.[46] [47] Also, overexpressed keratin proteins EGPR, ID4, TP53, and Ki67 have been linked to TNBC and BLBC as well.[48]

Second, look for genetic biomarkers that can determine the progression or prognosis, therapeutic response, and overall survivability to TNBC and BLBC. As an example[49], SFRP1 secretion which correlates with both an aggressive form of breast cancer, has

[44] Zhang, Shizhen, Zhen Wang, Weiwei Liu, Rui Lei, Jinlan Shan, Ling Li, and Xiaochen Wang. "Distinct Prognostic Values of S100 mRNA Expression in Breast Cancer." *Scientific Reports* 7, no. 1 (January 4, 2017).

[45] Sizemore, Gina M., Steven T. Sizemore and Darcie D. "GABA(A) Receptor Pi (GABRP) Stimulates Basal-like Breast Cancer Cell Migration through Activation of Extracellular-Regulated Kinase 1/2 (ERK1/2)." *Journal of Biological Chemistry* 289, no. 35 (July 10, 2014): 24102–13.

[46] Friedenson, Bernard. (2007). 1471-2407.

[47] Yadav, 252-263

[48] Ibid

[49] Huelsewig, 2014.

responded positively to neoadjuvant chemotherapy. Other examples include TM4SF1 in MDA-MB-231 cells, FOXC1, CXRC4, microRNA-761, ARPC2, and SCD1 all of which are overexpressed in aggressive tumors, and those with a higher likelihood of cellular migration resulting in metastasis, and low survival.[50]

Third, identify genetic biomarkers responsible for the development and progression of TNBC and BLBC, which can also lead to the development of targeted therapeutics. One notable example[51] is SCD1 inhibition that has been reported to reduce the proliferation and survival of cancer cells. Another is down-regulation and depletion of TRIM28, which has shown reduction in tumor growth.[52] In vivo studies also showed that expression of AMD3100 or siRNA in MDA-MB-231 could inhibit CXRC4 by under-expressing FOXC, thus controlling proliferation and metastasis.[53]

Gene Expression Profiling

DNA microarray technology has been commonly used in many biological purposes such as gene expression analysis, environmental monitoring, disease characterization. Its application in gene expression profiling is based on a multiplex technology used to simultaneously access thousands of genes and identify genes who are differentially expressed in response to "pathogens" by comparing gene expression between infected and uninfected cells or tissues.[54] A DNA microarray chip consist of an arrayed series of microscopic spots with immobilized gene-specific DNA oligonucleotides probes. The

[50] Fan, Chunni, Ning Liu, Dan Zheng, Jianshi Du, and Keren Wang, 6755–64.
[51] Hosokawa, Yuko, Noritaka Masaki and Shiro Takei, 2017.
[52] Czerwińska, Patrycja, Parantu K. Shah and Katarzyna Tomczak et al, 2016.
[53] Pan, Hongchao, Zhilan Peng, Jiediao Lin and Xiaosha Ren, 3794–3804.
[54] J Mehta. Gene expression analysis in breast cancer. 2010

hybridization of the fluorophore-labeled target onto the probe is usually detected and quantified to determine relative abundance of target.[55]

In Affymetrix microarrays, the probes are attached to the substrates by a covalent bond through a photolithographic process. Each GeneChip contains around 1,000,000 probe sets that are intended to measure expression for a specific mRNA. Each probe set consists of probe pairs selected from the target sequence which is derived from one or more mRNA sequences. The first pair is a perfect match (PM), and the other is mismatch (MM) at the center. This allow the quantitation and subtraction of nonspecific signals cross-hybridization.[56] Each gene or transcript consists of 11 probe pairs on the GeneChip, each name of which has a suffix consisting of the last three or four characters of its name that describes their ability to bind different genes, splice variants, or their uniqueness as it shown below[57]:

- "_at" hybridizes to unique anti-sense transcript of the gene.
- "_a_at" all probes cross-hybridize to the same set of sequences from the same gene family.
- "_s_at" all probes cross-hybridize to the same set of sequences, but these sequences are not from the same gene family.
- "_x_at" at least one probe cross-hybridize with other target sequences.

Microarray technology has been used, since its early development, to identify gene expression profiles of clinical breast cancer cell lines and specimens. Some of the breast

[55] J Mehta. Gene expression analysis in breast cancer. 2010
[56] Ibid
[57] Ibid

cancer sub-groups that have been identified using this technology are the basal sub-type, and normal-like, luminal A, luminal B, and ERBB2 over-expressing.[58]

Data Mining in Gene Expression

Data mining can be used with gene expression data to discover patterns and develop knowledge from biological databases using information technology and computational techniques. Data mining is an automated data analysis process to find relationships among data elements. Many of these relationships are not obvious due to the large amount of the data. Therefore, the researches and scientists can use the data mining techniques to extract useful information and create knowledge from data to identify correlations between elements.[59]

The most common types of microarray data analysis in data mining includes gene selection, clustering, and classification[60]. The method of analysis can be determined depending on the nature of the data and the desired knowledge, using either a descriptive or predictive model. A descriptive model is used to identify patterns and relationships among the data, while a predictive model is used to predict the data using existing patterns.

There are several data mining software that can be used to perform data mining, such as SAS Enterprise Miner, S-Plus, SPSS, IBM Intelligent Miner, SGI MineSet, Microsoft SQL Server 2000, and Inxight VizServer. However, some biological data mining tools have been developed, such as Statistics for Microarray Analysis, Affymetrix Data Mining,

[58] T Sorlie, Perou and R Tibshirani. Gene expression patterens of breast caracinomas distinguish tumor subclass with clinical implication. pp10869-10874

[59] G Tzanis. Biological data mining (Scientific Programming)

[60] Piatetshy-Shapiro and Tamayo. Microarray Data Mining: Facing The Challenges. SIGKDD Explorations. 1-5

GeneSpring, VectorNTI, Spot Fire, and COMPASS.[61] However, in our study SAS Enterprise Miner software will be used to perform data mining analysis.

SAS Enterprise Miner

SAS Enterprise Miner software is an advanced tool to help users to perform data mining by developing either descriptive or predictive models. SAS software provides a variety of data mining tasks, including decision tree, neural networks, link analysis, and linear and logistic regression.[62]

Classification

In this study, classification task will be used to perform data mining, which is a process of learning a function that classifies a data element into two or several classes. Classification is mostly used in microarray analysis to distinguish diseases or identify the most efficient treatment for given genetic signature or predict outcomes by performing a predictive model based on known gene expression patterns.[63]

The most popular microarray data mining methods for classification include Support Vector Machine (SVMs), Neural Networks, K-nearest neighbors, classification/Decision trees, voted classification, weighted gene voting, and Bayesian classification.[64] However, in this study, Artificial Neural Network, Logistic Regression, Decision Tree, Least Angle

[61] Han. "How can data mining help bio-data analysis"? (Department of Computer Science). 1-2

[62] SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA. https://www.sas.com/en_sa/home.html

[63] G Tzanis. Biological data mining (Scientific Programming)

[64] ibid

Regression, Bayesian classifier, Gradient Boosting, SVMs, and Random Forrest models were investigated.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are known as “massively parallel processors, which tend to preserve experimental knowledge and enable their further use”.[65] This model was created based on learning the processes of the neurological function of the brain and the cognitive system.[66] ANNs can provide an extreme complexity of non-linear functions and predict new observations. The advantage of the parallel computing environment of ANNs allows users to improve the predictive power algorithm.[67] Applying neural networks enables users to build models with significantly more lift by allowing more runs to enhance predictive power incrementally. Additional features of this model include smart defaults for most neural network parameters, automatic selection of a validation data group, and automatic standardization of input data and targeted variables.[68]

Logistic Regression

Logistic regression is a statistical method used to analyze database to classify cases into the most likely category based on one or more independent variables. In this study, linear regression cannot be applied since the response variable is discrete.

[65] Hajek P. Municipal credit rating modelling by neural networks, Decision Support Systems. 108-118

[66] Palwal M, Kumar U. Neural networks and statistical techniques: A review of applications Expert Systems with Application. (2009), 2-17

[67] ibid

[68] Zolbanin H.M. Predicting overall survivability in comorbidity of cancers: A data mining approach. (Decision Support Systems) 150-161

Therefore, the regression will predict the odds of its occurrence into two categories instead of predicting the estimation point. Some of the advantages of using Logistic Regression includes selecting variables and modeling capabilities for unordered multinomial data.[69]

Decision Tree

Decision tree is a classification algorithm starts with a single node, which branches into possible class predictions the size of the decision tree and classification accuracy are used to determine the quality of the model analysis. Decision tree uses mathematical algorithms to identify a variable, also to corresponding threshold for that identified variable, which is branches the input data into two or more subgroups. The process is repeated at each node until the tree fully construed. The spilt search algorithm uses corresponding threshold to maximize the homogeneity of the outcome subgroups. The most common mathematical algorithm that used to split the observations into multiple classes are entropy-based information gain, Gini index, and Chi-square test.[70] Finally, the complexity of the tree is optimized via pruning between training and validation sets. Advantages of Decision trees include ease of deployment, interpretation and visualization.

Random Forest

[69] Zolbanin H.M,150-161

[70] Ibid

Random Forest is a collection of multiple Decision trees. It is a supervised learning algorithm that draws random samples from the training dataset to grow the trees of the forest to the largest extent possible. Trees are trained in parallel and no pruning is carried out to reduce the size of the trees. In order to categorize new data, it is first inputted to each of the trees to generate classifications or votes based on the selected variables. The Random Forest then chooses the final classification outcome based on the votes scored among all the trees. Some of the advantages of using Random Forest includes high accuracy, ability to handle large volumes of multidimensional data, and effective imputation of missing data among others.[71]

Support Vector Machine

Support Vector Machine (SVM) is a class of machine learning algorithms that enables users to fit a discriminant function of features such as polynomial and sigmoid nonlinear kernels to separate one class from another. The nonlinear kernel transforms the input data to a high dimensional space such that the data space become separable using two parallel separating hyperplanes. The distance between the parallel hyperplanes is maximized to the extent in which an optimized classification model can be realized. An SVM model can also be used for outlier detection and regression. Some of the SVM's advantages include effective handling of unbalanced data and less complexity compared to other classifiers such as ANNs Previous research has reported that SVM can provide diagnosis ability with high accuracy in cancer prediction.[72]

[71] Zolbanin H.M, 150-161.

[72] Ibid

Least Angle Regression

Least Angle Regression (LARS) belongs to a family of generalized linear models designed to handle high-dimensional data. The algorithm uses a forward stepwise selection method to identify the optimal variable set; however, instead of adding variables at each step based on some pre-specified criteria such as adjusted R^2 or Akaike, it selects the variable that is most correlated with the target variable and then increases the estimated parameters in the least-squares direction until another variable has as much correlation with the target as the current one has. This selection process is repeated until none remain to be chosen. Some major advantages of LARS method are its abilities to handle high dimensional multi-collinear data and identify the best set of variables. [73]

Bayesian Classifier

Bayesian Classifiers are a family of simple probabilistic models that rely on Bayes' theorem to make class predictions given some data. In Bayesian machine learning, the input variables are assumed to be independent from each other. To classify a new observation, it simply estimates the probability that the given data point falls in a certain class and at the end, chooses the classification that has the highest probability. Some of the advantages of using a Bayesian Classifier model includes handling continuous and discrete data, making probabilistic predictions, and requiring less training data.

[73] Zolb Zolbanin H.M, 150-161.

Gradient Boosting

Gradient Boosting is a supervised machine learning algorithm for classification and regression applications. It is an ensemble of many prediction models using decision trees. Unlike Random Forest that uses random samples to build independent trees in parallel, Gradient Boosting builds trees one at a time in a sequential manner such that each tree is dependent on the residuals of the previous one. Gradient Boosting first draws a random sample (with replacement) from the original data, trains a decision tree, and tests its performance on the entire data. Then, the next random sample is drawn from the original dataset, which includes data points that were misclassified with the previous tree, and used that to build the second tree, and so on. This process is repeated until the error function does not change. While Gradient Boosting provides a very high predictive accuracy, it is less interpretable and prone to over-fitting due to its greater flexibility in fitting data.[74]

Data Pre-processing

Data pre-processing for data mining is a critical step to get better results. The data pre-processing is a process of cleaning the data from missing, out of range, or invalid values. It also provides several features such as understanding what the data represents, exploring variable statistics and distributions, performing appropriate transformations, and reducing the data among others. However, data pre-processing is time-consuming, but it is an

[74] Zolbanin H.M, 150-161.

important step to ensure the accuracy of the final results. Data pre-processing takes approximately up to 80% of the overall time of data mining.[75]

Measures for Performance Evaluation

There are three common performance measures used in binary classification models. The first is accuracy which determines the overall classification performance of the model; it calculates the percentage of correctly classified instances. Second, sensitivity which measures the proposition of positives that correctly identified. The third is specificity which measures the proposition of negatives that correctly identified. These performance measures can be obtained mathematically by the following expressions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.[76] Some studies use the misclassification rate to evaluate performance accuracy. However, the SAS Enterprise Miner software uses the misclassification rate in the validation group to rank the models depend on their performance accuracy.

Related Work

[75] Zolbanin H.M, 150-161.

[76] ibid.

A literature survey showed that there are many studies on the characterizing basal-like breast cancer using statistical analysis. However, we could only find a few studies related to characterizing BLBC using data mining approaches.

Rody et al.[77] conducted a research study regarding TNBC and BLBC by using a database generated on Affymetrix gene chips for 579 TNBC to perform unsupervised analysis to propose a definition of metagenes that differentiate molecular subset within TBNC without considering any clinical outcome. A single platform (Affymetrix U133a AND u133 Plus 2.0 chips) was used for data. However, 394 cases used for discovery, while 185 cases for validation. 16 metagenes expressions were correlated with survival and multivariate analysis, including pathological and routine clinical. Those metagenes includes basal-like phenotype, apocrine/androgen and cludin-low molecular subtypes, or reflected various non-neoplastic cell population, including blood, stroma, immune cells, adipocytes, inflammation and angiogenesis within the cancer.

In this study, Rody et al. observed a transparent bimodal distribution of basal-like metagene score within TNBC. Based on the bimodal distribution, a cutoff (0.0014) was driven to separate cases into low and high expressions groups by fitting two normal distributions as shown in Figure 1. As a result, 72.8% of TNBC were classified as BLBC in the discovery cases, while 69.7% of in validation cases.

In our study, we used the same TNBC database, and the cutoff value (0.0014) of BLBC to average the important variables. We defined relevant genes from the data as the average expression of high co-expressed genes groups without considering clinical outcomes.

[77] Rody, 2011

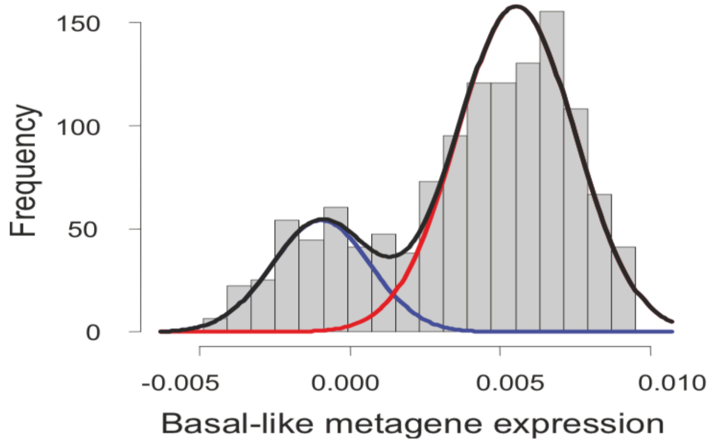

Figure 1. Distribution of the expression of basal-like metagene among TNBC.[78]

Methodology

It is widely known among data scientists that big data is composed of not only a large volume of data but also from several different sources, in various formats, from which greater insights can be gleaned. Therefore, a considerable amount of time and effort needs to be devoted on data management. Moreover, recent development in high performance analytical methods has improved our ability to extract meaningful insights from high dimensional data, which can be investigated using statistical analysis. In this section, we discuss how these important tasks are accomplished. We describe the data and research methodology used in this study in the following subsections. Our research methodology consists of four major phases: data acquisition, data integration, data preprocessing, and predictive modeling. The analytical methodology is depicted in Figure 2.

[78] Rody, 2011

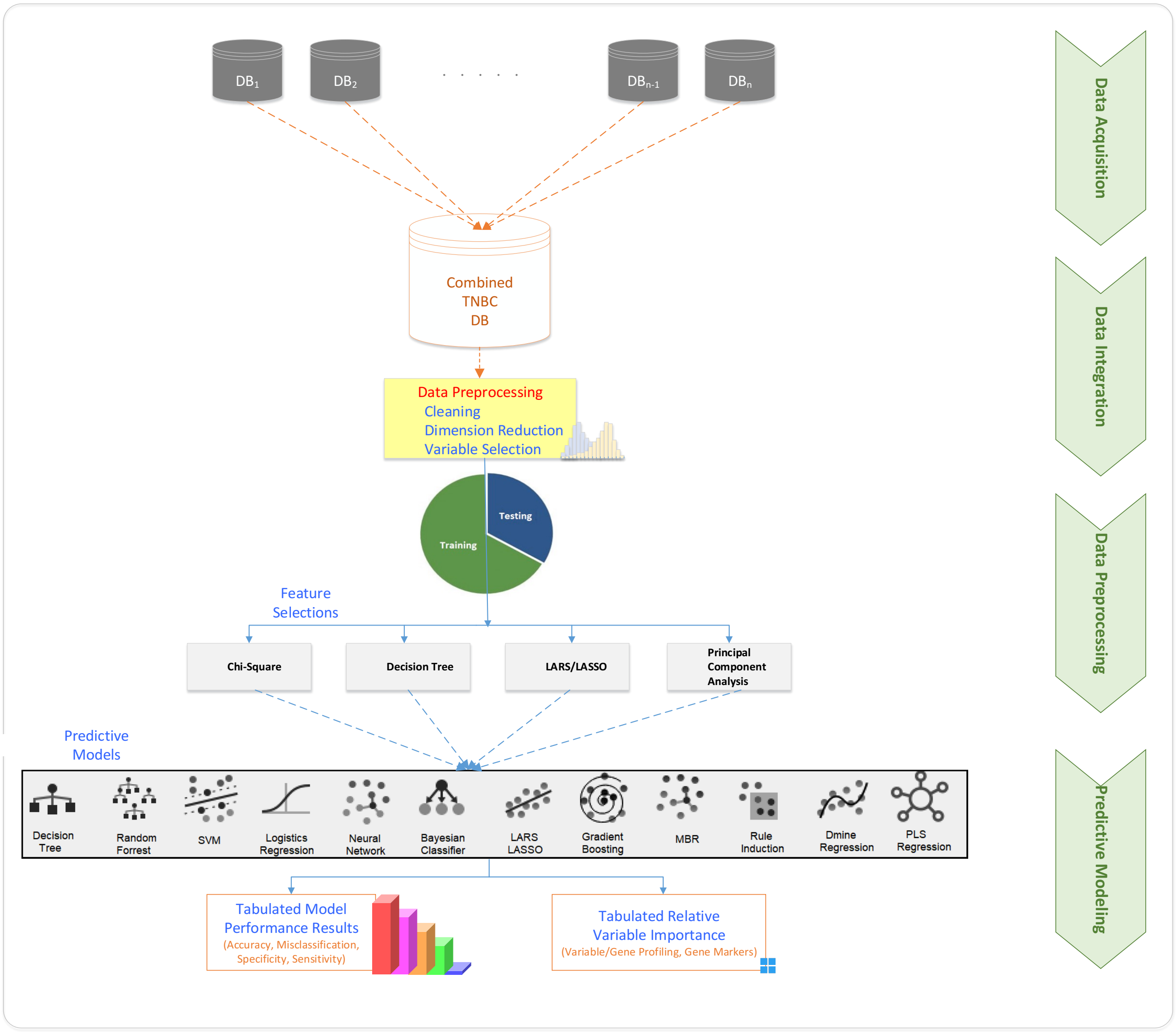

Figure 2. Research Methodology.

However, it is important to mention that other prognostic factors in BC such as age, histological grade, and tumor size were not considered. The grade has no significant regard of prognosis since most TNBC cases are high grade. Also age and tumor size factors are not considered since TNBC subtype is associated with younger age, so the impact of these two factors for prognosis in TNBC is not yet fully clear.[79]

[79] Rody, 2011

Data Acquisition and Integration

Microarray data generation is a very expensive process; therefore, collecting large data microarray is challenging and requires a substantial amount of resources. To build a large sample size for this study, it was necessary to pool several datasets from different laboratories.[80] We used multiple public datasets that were built according to the most widely microarray platform (Affymetrix U133A and U133 Plus 2.0 chips) and included only cases that were defined as triple negative based on the mRNA expression of ER, PgR, and HER2 as previously described.[81] We compiled a total of 3,488 publicly available breast cancer gene expression profiles from 28 individual datasets and extracted Affymetrix gene expression data for 579 TNBC cases.

Data Preprocessing

Data preprocessing is a critical step in data mining, which involves data cleaning, variable reduction and feature selection. It involves cleaning the data from missing, out of range, or invalid values. It also allows for a better understanding of what the data represents, to explore variable statistics and distributions, to perform appropriate transformations, and to reduce the data, among others. Although data preprocessing is time consuming, it is an important step in ensuring the accuracy of the results. Data preprocessing takes up to 80% of the overall time of data mining.

[80] Rody, 2011

[81] Sizemore, Gina M., Steven T. Sizemore, Darcie D. Seachrist, and Ruth A. Keri, 24102–13

With microarray data being high dimensional, characterized by many variables and few observations, it requires feature selection and dimension reduction techniques to remove genes that do not provide significant incremental information. In this study, we observed five missing genes in some of the expression datasets, so we excluded those genes from the analysis. Moreover, we applied various feature selection methods, such as chi-square, decision tree, Least Angle Regression/ Least Shrinkage and Selection Operator (LARS/LASSO), principle component analysis (PCA) and ensemble (multi-method) algorithms, to identify key variables (genes) that could explain the differences in the observations and could be used to simplify the analysis and prediction of BLBCs. It is recognized that different feature selection techniques may result in different sets of biomarkers, that is, different groups of genes highly correlated to a given condition; however, together, these results can be used to identify driving pathways in basal-like breast cancer.

Predictive Models

In this study, we used five different feature selection methods (i.e., Chi-square, tree, LARS, LASSO and ensemble), along with eight predictive models (i.e., Logistic Regression, Decision tree, Random Forest, Support Vector Machine, Neural Networks, LARS, Gradient Boosting, and Bayesian Classifier) in an empirical investigation to understand, characterize and predict BLBCs. The data was divided into 70% for training and 30% for validation. We performed supervised analysis to define a set of gene markers that distinguished molecular subsets within TNBCs. The 574 cases were divided into 394 for discovery and 185 for validation. The initial step was to build stratified datasets for this analysis. The second step involved applying various popular data mining techniques, including Decision trees, Regression analysis, Random Forest, Neural Network, Least Angle Regression, Bayesian Classifier, Gradient Boosting, and Support Vector Machine, to classify BLBC and non-BLBC cases and to identify structures in the molecular data of the targeted disease. Of all the cases evaluated, 394 were used for training and 185 for validation. More than 22,000 genes expression data were correlated with survival using multivariate analysis, including pathological and routine clinical data. Those metagenes included the basal-like phenotype, apocrine/androgen and claudin-low molecular subtypes, or reflected various non-neoplastic cell populations, including blood, stroma, immune cells, adipocytes, inflammation and angiogenesis within the cancer. In summary, 40 different predictive models were built to identify gene signatures and determine which genes contribute most to BLBC.

We used two types of prediction models in SAS enterprise Miner 12.3 software. The first type is partial data, which allows us to divide the data set for two groups, 70% for the data used for train, and 30% for validation. Thus, 579 TNBC cases divided into 394 cases for the discovery cohort and 185 cases for validation. High-performance data partition used train data for preliminary model fitting, whereas Validation data used to assess the adequacy of the fitted model.[82]

The other type is high-performance data mining (HPDM), which provides several advantages, including reductions of dimensions for structured inputs and perform unsupervised variable selection. The high-performance regression aims to predict the probability of a binary target acquiring an interest event of the assigned link function of

[82] SAS Institute Inc. 2011. *SAS Enterprise MinerTM High-Performance Data Mining Node Preference for SAS 9.3.* Cary, NC Institute INC.

one or more independent inputs. In our study, we used the cutoff value of 0.0014 from Rody et al. to average the crucial variables and assigned 1 to be BLBC class, and 0 for non-BLBC.

After implementing these two types, the following models utilized: Artificial Neural Network, Logistic Regression, Decision Tree, Least Angle Regression, Bayesian Classifier, Gradient Boosting, Support Vector Machine, and Random Forest along with these nodes.

Result

Performance Evaluation of the models

A summary of the model's performances on 22,000 gens of 579 TNBC cases used in this study is shown in Table 1 The table includes the misclassification, accuracy, sensitivity, and specificity rates for each classifier. According to the result, the neural network model shows the highest average accuracy compared to other methods. The Gradient Boosting and Logistics Regression models have very close values of accuracy, while the Decision tree has the lowest average accuracy.

Table 1. Models' performance Evaluation.

Feature Selection method: Chi-Square				
Method	Misclassification	Accuracy	Sensitivity	Specificity
Decision Tree	0.0455	0.9545	0.9922	0.8542
LARS	0.0511	0.9489	0.9766	0.8750
Neural Network	0.0455	0.9545	0.9688	0.9167
Logistics Regression	0.0398	0.9602	0.9688	0.9375
SVM	0.0230	0.9770	0.9921	0.9362
Gradient Boosting	0.0341	0.9659	0.9922	0.8958
Random Forrest	0.0341	0.9659	0.9922	0.8958
Bayesian Classifier	0.0520	0.9480	0.9680	0.8958
Feature Selection method: Decision Tree				
Method	Misclassification	Accuracy	Sensitivity	Specificity

Decision Tree	0.0517	0.9483	0.9921	0.8298
LARS	0.0345	0.9655	0.9764	0.9362
Neural Network	0.0556	0.9444	0.9398	0.9574
Logistics Regression	0.0287	0.9713	0.9843	0.9362
SVM	0.0230	0.9770	0.9921	0.9362
Gradient Boosting	0.0230	0.9770	0.9843	0.9574
Random Forrest	0.0402	0.9598	0.9921	0.8723
Bayesian Classifier	0.0575	0.9425	0.9528	0.9149
Feature Selection method: LARS/LASSO				
Method	Misclassification	Accuracy	Sensitivity	Specificity
Decision Tree	0.0625	0.9375	0.9922	0.7917
LARS	0.0511	0.9489	0.9766	0.8750
Neural Network	0.0118	0.9882	0.9918	0.9792
Logistics Regression	0.0345	0.9655	0.9764	0.9362
SVM	0.0455	0.9545	0.9844	0.8750
Gradient Boosting	0.0172	0.9828	0.9921	0.9574
Random Forrest	0.0455	0.9545	0.9844	0.8750
Bayesian Classifier	0.0625	0.9375	0.9531	0.8958
Feature Selection method: Principal Component Analysis				
Method	Misclassification	Accuracy	Sensitivity	Specificity
Decision Tree	0.1207	0.8793	0.9606	0.6596
LARS	0.0460	0.9540	0.9764	0.8936
Neural Network	0.0345	0.9655	0.9764	0.9362
Logistics Regression	0.0402	0.9598	0.9764	0.9149
SVM	0.0517	0.9483	0.9843	0.8511
Gradient Boosting	0.0805	0.9195	0.9685	0.7872
Random Forrest	0.0707	0.9293	0.9781	0.7872
Bayesian Classifier	0.1034	0.8966	0.9370	0.7872
Feature Selection method: Multi-method				
Method	Misclassification	Accuracy	Sensitivity	Specificity
Decision Tree	0.0862	0.9138	0.9606	0.7872
LARS	0.0230	0.9770	0.9921	0.9362
Neural Network	0.0172	0.9828	0.9843	0.9787
Logistics Regression	0.0287	0.9713	0.9843	0.9362
SVM	0.0287	0.9713	0.9843	0.9362
Gradient Boosting	0.0172	0.9828	0.9921	0.9574
Random Forrest	0.0230	0.9770	0.9921	0.9362

Bayesian Classifier	0.0230	0.9770	0.9843	0.9574

The results indicate that most predictive models gained prediction accuracy from a multi-method feature selection approach, compared to each individual approach. We found no evidence that a certain feature selection method is particularly well suited for use in combination with a specific predictive model. However, Decision tree, Gradient Boosting, Random Forest, and Bayesian Classifier did not gain much prediction accuracy from one principal component compared to another.

The feature selection analysis revealed over 500 genes, which appear to be associated with BLBC. Table 2 summarizes the top 40 genes from pathways, which are associated with BLBC. These genes are sorted according to the number of times they were selected by the feature selection algorithms as input for predictive models..

Table 2. Top 40 Genes Associated to BLBC.

Gene	Gene Description	Count
_205044_at	Gamma-aminobutyric acid (GABA), A receptor, pi(GABRP)	14
_220425_x_at	ROPN1B	14
_204855_at	Serpin peptidase inhibitor, clade B (ovalbumin), member 5 (SERPINB5)	11
_213260_at	Forkhead box C1 (FOXC1)	11
_205157_s_at	Keratin 17, type I (KRT17)	10
_209800_at	keratin 16, type I (KRT16)	10
_202037_s_at	Secreted frizzled-related protein 1 (SFRP1)	9
_206560_s_at	Melanoma inhibitory activity (MIA)	9
_209387_s_at	Transmembrane 4 L six family member 1 (TM4SF1)	9
_219768_at	V-set domain containing T cell activation inhibitor 1 (VTCN1)	8
_209504_s_at	Pleckstrin homology domain containing, family B (evectins) member 1 (PLEKHB1)	7
_211682_x_at	UDP glucuronosyltransferase 2 family, polypeptide B28 (UGT2B28)	7
_212236_x_at	JUP	7
_60474_at	Fermitin family member 1 (FERMT1)	7
_201820_at	Keratin 5, type II (KER5)	6
_208998_at	Uncoupling protein 2 (mitochondrial, proton carrier) (UCP2)	6

_209126_x_at	Keratin 6B, type II (KRT6B)	6
_210473_s_at	Adhesion G protein-coupled receptor A3 (ADGRA3)	6
_202036_s_at	Secreted frizzled-related protein 1 (SFRP1)	5
_202504_at	Tripartite motif-containing 29 (TRIM 29)	5
_204751_x_at	DSC2	4
_213680_at	Keratin 6B, type II (KRT6B)	4
_217901_at	Desmoglein 2 (DSG2)	4
_218868_at	ARP3 actin-related protein 3 homolog B (yeast) (ACTR3B)	4
_200832_s_at	Stearoyl-CoA desaturase (delta-9-desaturase) (SCD)	3
_201485_s_at	Reticulocalbin 2, EF-hand calcium binding domain (RCN2)	3
_202342_s_at	Tripartite motif containing 2 (TRIM 2)	3
_203058_s_at	3'-phosphoadenosine 5'-phosphosulfate synthase 2 (PAPSS2)	3
_204268_at	S100 calcium binding protein A2 (S100A2)	3
_205265_s_at	SPEG complex locus (SPEG)	3
_207397_s_at	Homeobox D13 (HOXD13)	3
_208063_s_at	Calpain 9 (CAPN9)	3
_209351_at	Keratin 14, type I (KRT14)	3
_209791_at	Peptidyl arginine deiminase, type II (PADI2)	3
_209842_at	SRY (sex determining region Y)-box 10 (SOX10)	3
_210074_at	Cathepsin V (CTSV)	3
_212147_at	Nonsense mediated mRNA decay factor (SMGS)	3
_214598_at	Claudin 8 (CLDN8)	3
_219301_s_at	Contactin associated protein-like 2 (CNTNAP2)	3
_219795_at	Solute carrier family 6 (amino acid transporter), member 14 (SLC6A14)	3

Neural Network (ANNs)

Neural Network (ANNs) considered to be one of the powerful methods to analyze the data with high accuracy. In this study, this model identified 12 important genes that correlate with BLBC. Figure 1 shows the equation’s line between those gens and BLBC class. Those gens are Transglutaminase 2(TGM2), Discs, large homolog 5 (Drosophila)(DLG5), Cytochrome b5 reductase 1(CYB5R1), Desmocollin 2(DSC2), Transmembrane protein 5(TMEM5), GDP-mannose 4,6-dehydratase(GMDS), Gamma-aminobutyric acid (GABA) A receptor, pi(GABRP), Phospholipase A2, group IB (pancreas)(PLA2G1B), Junction plakoglobin(JUP), Chromosome 19 open reading frame

73(C19orf73), Rhophilin associated tail protein 1B (ROPN1B), and Nuclear factor I/X (CCAAT-binding transcription factor) (NFIX).

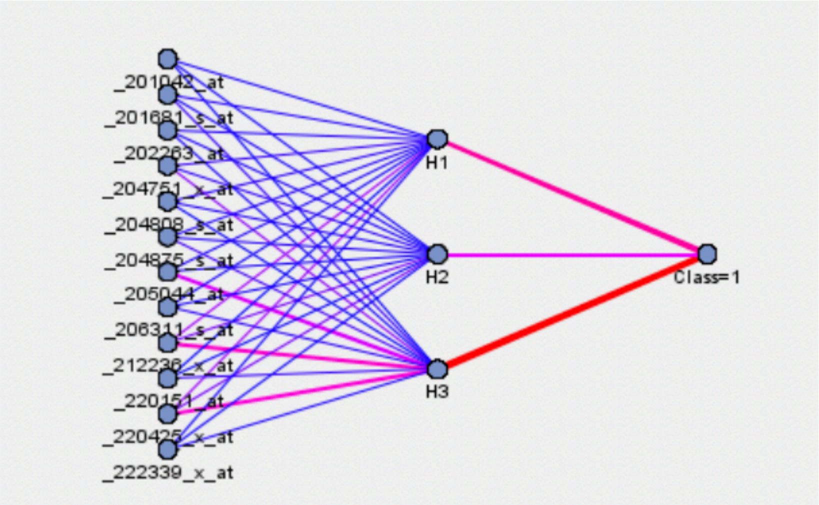

Figure 3. Neural Network Model (ANNs).

In Figure 3, the small blue circles on the far left of the link graph represent all variables input (genes), which have correlation with BLBC. The target variable placed on the far right of the link graph; which is in this case represents class=1 (the BLBC subtype). H1, H2, and H3 are the hidden layers. The color and the width of the linked lines indicate how secure the connection is of that particular line; the thinner, blue lines represent a smaller value of the weight of that connection, and the thicker red line indicates a substantial magnitude value of the link connection.[83]

In Figure 3, chromosome 19 open reading frame 73(C19orf73), phospholipase A2, group IB (pancreas)(PLA2G1B), and GDP-mannose 4,6-dehydratase(GMDS) have pink and thicker lines, which means that the magnitude of the weight of the connection is

[83] SAS Institute Inc. 2011. *SAS Enterprise MinerTM High-Performance Data Mining Node Preference for SAS 9.3.* Cary, NC Institute INC.

significant. These genes linked to hidden "layer 3" (H3); H3 linked to class=1 (BLBC) by the red, thicker line, which shows a strong connection. As a result, C19orf73, PLA2G1B, and GMDS have a higher correlation to BLBC, followed by CYB5R1 and JUP.

These 12 genes are known to be involved in various aspects of TNBC's pathogenesis. For instance, TGM2 is involved in TNBC epithelial-to-mesenchymal transition (EMT), which promotes their migratory and invasive properties, and controls their chemoresistance and immune escape.[84] In addition, TGM2 expression is frequently up-regulated during inflammation and wounding. Emerging evidence indicates that TGM2 expression is aberrantly up-regulated in multiple cancer cell types, particularly those selected for resistance to chemotherapy and radiation therapy and those isolated from metastatic site.[85]

Loss of DLG5 promotes TNBC cell proliferation by inhibiting the Hippo signaling pathway, increasing nuclear YAP expression, and inducing EMT.[86] DLG5 plays important roles in epithelial cell polarity maintenance, precursor cell division, cell proliferation, cell migration and invasion, and transmission of extracellular signals to the membrane and cytoskeleton. Failure in establishment and maintenance of epithelial cell polarity contributes to tumorigenesis. Loss of expression and function of cell polarity proteins is

[84] W He, Z Sun, Z Liu. (2015) Silencing of TGM2 reverses epithelial to mesenchymal transition and modulates the chemosensitivity of breast cancer to docetaxel. *Exp Ther Med, 10*(4), 1413-1418

[85] N Agnihotri, S Kumar, and K Mehta. (2013). Tissue transglutaminase as a central mediator in inflammation-induced progression of breast cancer. *Breast Cancer Research, 15*(1), 202.

[86] Liu, J., Li, J., Li, P., Wang, Y., Liang, Z., Jiang, Y., . . . Chen, H. (2017). Loss of DLG5 promotes breast cancer malignancy by inhibiting the Hippo signaling pathway. *Scientific reports, 7*, 42125.

directly related to epithelial cell polarity maintenance.[87] Another gene whose expression correlates with EMT is CYB5R1, which is a widely expressed oxidoreductase involved in oxidative stress reactions and drug metabolism. Although its specific role in cancer progression is still not clear, its transcriptional level expression strongly correlates with EMT in colorectal cancer.[88]

The DSC2 protein is a major component of desmosomes, which provide strength and stability to tissues. This protein has been shown to be highly expressed in TNBCs, being able to significantly predict patient survival, and suggesting their role in the aggressiveness seen in these tumors.[89] TMEM5 is a type II transmembrane protein, thought to be a glycosyltransferase involved in the glycosylation of dystroglycan, which is part of a complex that links the extracellular matrix to the cytoskeleton. Aberrant glycosylation leads to the disruption of this link thus favoring migration and invasiveness seen in many tumors.[90] TMEM5 is significantly over-expressed in BRCA1-mutated breast cancer cells,[91] with these type of mutations occurring in TNBC more frequently than in the general population,[92] [93]. GMDS is involved in the process of cellular fucosylation of glycoproteins,

[87] Liu, J., Li, J., Ren, Y., & Liu, P. (2014). DLG5 in cell polarity maintenance and cancer development. *International journal of biological sciences, 10*(5), 543.

[88] Woischke, C., Blaj, C., Schmidt, E. M. (2016). CYB5R1 links epithelial-mesenchymal transition and poor prognosis in colorectal cancer. *Oncotarget, 7* 31350-31360.

[89] Hill, J. J., Tremblay, T. L., Fauteux, F., Li, J.(2015). Glycoproteomic comparison of clinical triple-negative and luminal breast tumors. *J Proteome Res, 14*(3), 1376-1388.

[90] Palmieri, V., Bozzi, M., Signorino, G., Papi, M.(2017). alpha-Dystroglycan hypoglycosylation affects cell migration by influencing beta-dystroglycan membrane clustering and filopodia length: A multiscale confocal microscopy analysis. *Biochim Biophys Acta Mol Basis Dis, 1863*(9), 2182-2191.

[91] Privat, M., Rudewicz, J., Sonnier, N. (2018). Antioxydation And Cell Migration Genes Are Identified as Potential Therapeutic Targets in Basal-Like and BRCA1 Mutated Breast Cancer Cell Lines. *Int J Med Sci, 15*(1), 46-58.

[92] Peshkin, B. N., Alabek, M. L., & Isaacs, C. (2010). BRCA1/2 mutations and triple negative breast cancers. *Breast Dis, 32*(1-2), 25-33.

[93] Chen, H., Wu, J., Zhang, Z., Tang, Y., Li, X., Liu, S., . . . Li, X. (2018). Association Between BRCA Status and Triple-Negative Breast Cancer: A Meta-Analysis. *Front Pharmacol, 9*, 909.

which involved in the functional regulation of adhesion molecules and growth factor receptors, with high levels of fucusylation being reported in various types of cancer.[94] This has been associated with TNBC and EMT, making GMDS a potential player in this process.[95]

Expression of GABRP is shown to be associated with the BLBC/TN subtype, and herein, we reveal its expression also correlates with metastases to the brain and poorer patient outcome.[96] PLA2G1B are esterases that preferentially cleave glycerophospholipids into biologically active fatty acids and lysophospholipids, and are differentially expressed in breast cancer.[97] These active lipids have biological functions relevant to cancer progression and each can be further metabolized into additional functional biomolecules.[98] These active lipids modulate cellular differentiation, proliferation, apoptosis and senescence, whose dysregulation can result in the uncontrolled growth and metastasis seen in tumors.

JUP is a cell adhesion protein, was recently reported as a determinant of circulating tumor cells types, single or clustered. This protein could be functioning as a double-edge sword, since loss of its expression leads to increased motility of epithelial cells, thereby promoting EMT and further metastasis. However, studies also show that JUP can function as an oncogene, with high expression of JUP resulting in clustered tumor cells in circulation with high metastatic potential in breast cancer and shortened patient survival. In addition, JUP

[94] Miyoshi, E., Moriwaki, K., & Nakagawa, T. (2008). Biological function of fucosylation in cancer biology. *J Biochem, 143*(6), 725-729.

[95] Listinsky, J. J., Siegal, G. P., & Listinsky, C. M. (2011). The emerging importance of alpha-L-fucose in human breast cancer: a review. *Am J Transl Res, 3*, 292-322.

[96] Sizemore, Gina M., Steven T. Sizemore and Darcie D., 24102-13.

[97] Yamashita, S., Ogawa, M., Sakamoto, K., Abe, T., Arakawa, H., & Yamashita, J. (1994). Elevation of serum group II phospholipase A2 levels in patients with advanced cancer. *Clin Chim Acta, 228*(2), 91-99.

[98] Scott, K. F., Sajinovic, M., Hein, J., Nixdorf, S., Galettis, P., Liauw, W., . . . Russell, P. J. (2010). Emerging roles for phospholipase A2 enzymes in cancer. *Biochimie, 92*(6), 601-610.

may be a potential prognostic biomarker that can be exploited to develop as a therapeutic target for breast cancer.[99] Although C19orf73 is a hypothetical protein that has not been characterized, a search of the GEO Profiles database[100] revealed that it is overexpressed in TNBC (GEO accession GDS4069.[101]

Ropporin is a sperm-specific protein and is associated with sperm motility. Its expression was also found in motile cilia helping them to move in one direction in a synchronized pattern. Ropporin (ROPN1 and ROPN1B) was identified as differentially-expressed in several gene lists commonly associated with bad prognosis in our breast cancer investigation.[102] The Nuclear Factor I (NFI) family of site-specific DNA binding proteins functions in adenoviral DNA replication and in the regulation of transcription of a large variety of cellular and viral genes. This family is comprised of four genes in vertebrates (NFIA, NFIB, *NFIC* and *NFIX*), whose encoded proteins interact with DNA as homo- or hetero-dimers. They bind to the palindromic sequence TTGGC(N5)GCCAA with high affinity, resulting in transcriptional activation or repression, depending on the cellular context and regulatory region . Binding sites for these factors have been identified in promoter, enhancer and silencer regions of a plethora of genes expressed in almost every

[99] L. Lu, H. Zeng, X. Gu and W. Ma, 491-500.

[100] Barrett, T., & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol, 411*, 352-369.

[101] Yang, L., Wu, X., Wang, Y., Zhang, K., Wu, J., Yuan, Y. C., . . . Yen, Y. (2011). FZD7 has a critical role in cell proliferation in triple negative breast cancer. *Oncogene, 30*(43), 4437-4446.

[102] J Mehta. Gene expression analysis in breast cancer. 2010.

organ and tissue.[103] It's been found that expression is increased in TNBC across the datasets [104] in a study but NFIX needs further studies.

Logistic regression

Logistic regression is another technique that we used in this analysis. In figure 4, the chart shows the relative importance to BLBC for 12 genes sorted in descending order. The horizontal axis is the correlated genes, while the vertical axis shows the value of the correlation range from 0-1 since the binary targets have two levels, where 1 represents the essential variables.

According to the results, gamma-aminobutyric acid (GABA) A receptor, pi (GABRP), has the highest correlation with BLBC based on the relative importance value; followed by JUP, ROPN1B, DSC2, TMEM5, GMDS, NFIX, C19orf73, DLG5, CYB5R, PLA2G1B, and TGM2, respectively.

[103] Becker-Santos, D. D., Lonergan, K. M., Gronostajski, R. M., & Lam, W. L. (2017). Nuclear factor I/B: a master regulator of cell differentiation with paradoxical roles in cancer. *EBioMedicine, 22*, 2-9.

[104] Han, W., Jung, E. M., Cho, J., Lee, J. W., Hwang, K. T., Yang, S. J., . . . Park, I. A. (2008). DNA copy number alterations and expression of relevant genes in triple-negative breast cancer. *Genes, Chromosomes and Cancer, 47*(6), 490-499.

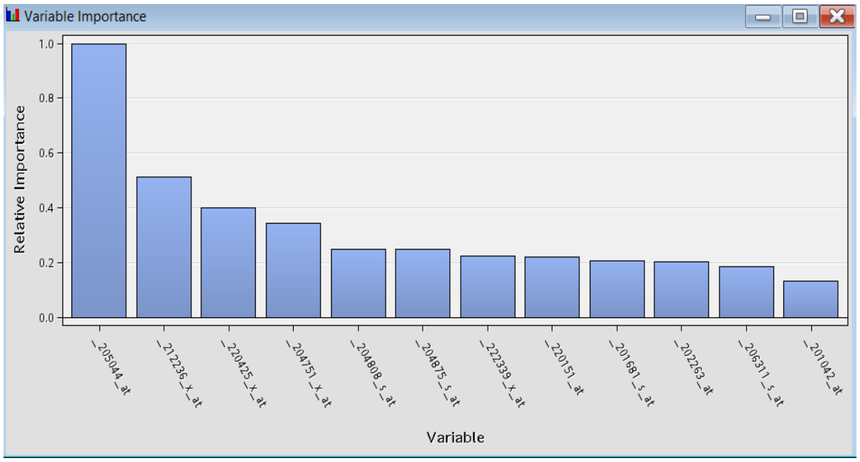

Figure 4. Logistic regression model.

Decision Tree

The decision tree model shows gene expressions in ascending order regard on the correlation to BLBC subtype. Each node includes some general properties such as the node Id, statistic information for both the train and validation group. As mentioned earlier, 1 represents the BLBC class and 0 for the non-BLBC class.

In figure 5, Node 1 represents gamma-aminobutyric acid (GABA) A receptor, pi(GABRP) as the highest correlated gene to BLBC; reflecting that a 72.73% of TNBC classified as BLBC in the validation group, and 27.27% are non-BLBC. On the other hand,72.95% are BLBC of the training group, and 27.05% are non-BLBC. However, two split nodes resulted according to the rule of "if values are less or more than the cutoff value (0.0014)". The corresponding genes are given in Table 3. However, based on a research was done to investigate the impact of gene expression on TNBC, all these identified genes were up-regulated in TNBC (more details are discussed in discussion section). Therefore, it can not be concluded if any up or down regulation is associated with TNBC from discussion model.

Table 3. Gene importance by Decision Tree model.

Gene	Importance	Description
GABRP	1	gamma-aminobutyric acid (GABA) A receptor, pi.
JUP	0.5097	junction plakoglobin.
ROPN1B	0.4033	rhophilin associated tail protein 1B.
DSC2	0.3393	desmocollin 2.
TMEM5	0.2383	transmembrane protein 5.

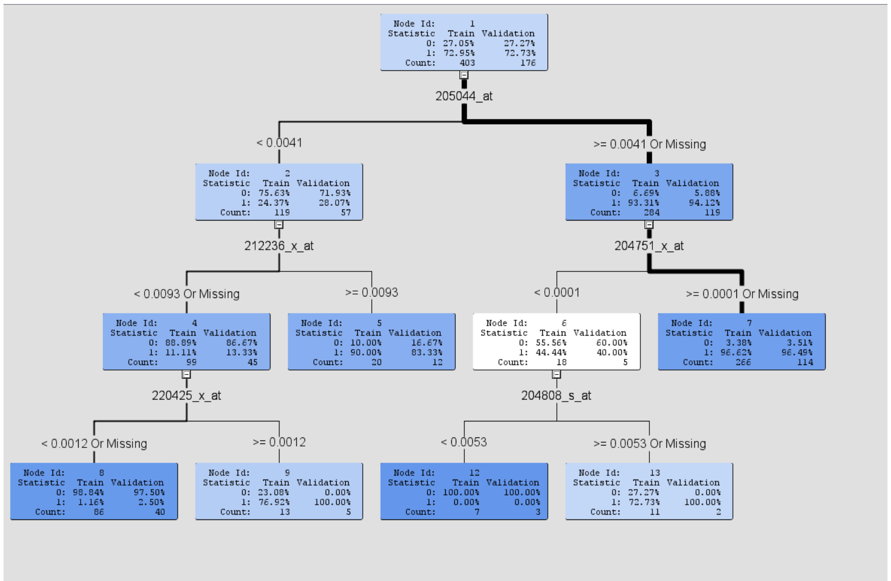

Figure 5. Decision Tree model.

Discussion tree algorithm, started single node by classifying GABRP gene as most important gene with score 1 out of 1 as shown in table 3. Then it branches to other possible classification based on the average importance value. Therefore, if it less than 0.0041 the second important gene is JUP gene as single node 2, from node 2 at average of 0.0093 another node branched to predict another relevant gene which is ROPN1B. at each node the statistical information is provided as mentioned before.

Random Forest

Random Forest model targeted 10 genes. Table 4 includes the number of splitting rules for each gene. However, As the rules splits more, the importance of the gene increases. Some genes have the same number of splitting rules, which means that those genes have the same level of importance.

Table 4. Random Forest with variable selection enabled.

Variable Name	Number of splitting rules	Gene
_202504_at	15.0	**TRIM 29**
_205157_s_at	12.0	KRT17
_202342_s_at	9.0	RCN2
_204855_at	9.0	SERPINB5
_206560_s_at	9.0	MIA
_214404_x_at	8.0	SPDEF
_219615_s_at	8.0	KCNK5
_202431_s_at	7.0	MYC
_205044_at	7.0	GABRP
_209504_s_at	6.0	PLEKHB1

Table 4 ranked the most important gene as following: TRIM2, Keratin 17, type I (KRT17), Reticulocalbin 2, EF-hand calcium binding domain (RCN2), Serpin peptidase inhibitor, clade B (ovalbumin), member 5 (SERPINB5), Melanoma inhibitory activity (MIA), SAM pointed domain-containing Ets transcription factor (SPDEF), Potassium channel subfamily K member 5 (KCNK5), Proto-Oncogene, BHLH transcription factor (MYC), GABRP, then Pleckstrin homology domain containing, family B (evectins) member 1 (PLEKHB1).

Least Angle Regression

Figure 6 represents Least angle regression model; the method estimates the correlation between the gen and BLBC. The blue bars represent a positive correlation with BLBC, so the highest gene expression estimates, the highest chance to be classified as BLBC. The red bars represent a negative correlation with BLBC. However, UDP glucuronosyltransferase 2 family, polypeptide B2(UGT2B28), and cytoplasmic linker associated protein 1(LASP1) have a negative correlation with BLBC.

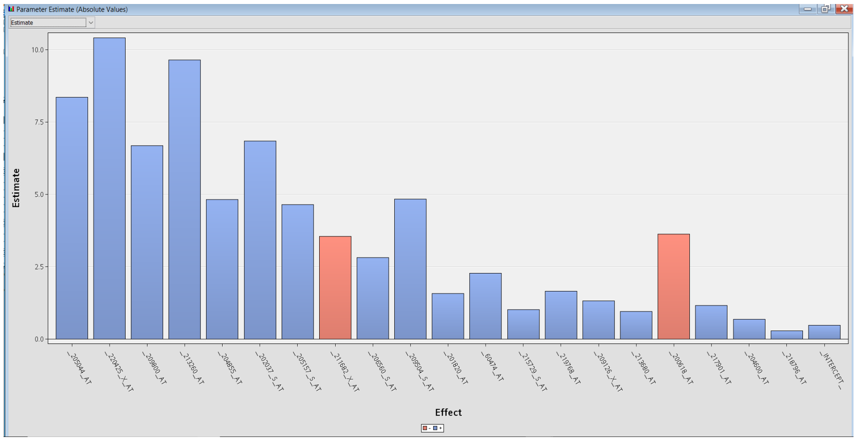

Figure 6 Least angle regression model.

In figure 6, the most critical genes based on the estimated value, are sorted in ascending order as following: ROPN1B, Forkhead box C1 (FOXC1), GABRP, Secreted frizzled-related protein 1 (SFRP1), keratin 16, type I (KRT16), SERPINB5, PLEKHB1, KRT17, MIA, Fermitin family member 1 (FERMT1), V-set domain containing T cell activation inhibitor 1 (VTCN1), Keratin 5, type II (KRT5), Keratin 6B, type II (KRT6B), Desmoglein 2 (DSG2), Vestigial like family member 1(VGLL1), following by EPH receptor B3(EPHB3).

Bayesian classifier

Bayesian Classifier is another method used to analyze the data. This model shows the genes most relevant to BLBC. The result of this model is represented in Table 5, where the most important 12 genes are presented.

Table 5. Gene Ranking (importance) using Bayesian Classifier.

Variable	Name of the gene	Order	Score
_220425_x_at	ROPN1B	1.0	-149.8388379843322
_205044_at	GABRP	2.0	-152.06028382414326
_212236_x_at	JUP	3.0	-178.30128029698253
_204751_x_at	DSC2	4.0	-197.10605719691765
_204875_s_at	GMDS	5.0	-250.02823043246084
_202263_at	CYB5R1	6.0	-252.2041273457195
_222339_x_at	NFIX	7.0	-255.19927200917277
_201042_at	TGM2	8.0	-256.2482325776195
_220151_at	C19orf73	9.0	-257.503065161863
_201681_s_at	DLG5	10.0	-260.21558211552104
_206311_s_at	PLA2G1B	11.0	-261.0894822749076
_204808_s_at	TMEM5	12.0	-261.13372976549573

Gradient Boosting

Gradient Boosting is the model with the highest accuracy performance in this study (97.15%). This model shows 12 most important genes, which have a high correlation with BLBC, as shown in Table 6. According to the results, ROPN1B is the most important gene expression, while GMDS in the least important among those 12 genes.

Table 6. Correlated genes with gen selection enabled.

NAME	GENE	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
_220425_x_at	ROPN1B	220425_x_at	24	1	1	1
_205044_at	GABRP	205044_at	13	0.976866	0.843971	0.863958
_212236_x_at	JUP	212236_x_at	20	0.6846	0.570677	0.833592
_204751_x_at	DSC2	204751_x_at	7	0.403207	0.437559	1.085197
_201042_at	TGM2	201042_at	3	0.093932	0	0
_201681_s_at	DLG5	201681_s_at	2	0.093776	0	0
_202263_at	CYB5R1	202263_at	2	0.084822	0	0
_204808_s_at	TMEM5	204808_s_at	2	0.062419	0	0
_220151_at	C19orf73	220151_at	1	0.061109	0.060705	0.993387
_222339_x_at	NFIX	222339_x_at	2	0.059511	0	0
_206311_s_at	PLA2G1B	206311_s_at	1	0.032846	0	0
_204875_s_at	GMDS	204875_s_at	0	0	0	

Summary

Table 7 shows all the relevant genes to BLBC with the model's names since some genes resulted as having association with BLBC in more than two models. The table helped to identify the most important genes. Since each model has a unique algorithm to classify those genes, the probability of the identifying the most correlated genes will high.

Table 7. Gene's list with models.

Gene	**Models**

ROPN1B	Neural Network, Logistic Regression, Decision Tree, Least Angle Regression, Bayesian classifier, Gradient Boosting.
GABRP	Neural Network, Logistic Regression, Decision Tree, Random Forest, Least Angle Regression, Bayesian classifier, Gradient Boosting.
TGM2	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
JUP	Neural Network, Logistic Regression, Decision Tree, Bayesian classifier, Gradient Boosting.
DSC2	Neural Network, Logistic Regression, Decision Tree, Bayesian classifier, Gradient Boosting.
TMEM5	Neural Network, Logistic Regression, Decision Tree, Bayesian classifier, Gradient Boosting.
GMDS	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
CYB5R1	Neural Network, Bayesian classifier, Gradient Boosting.
DLG5	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
PLA2G1B	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
C19orf73	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
NFIX	Neural Network, Logistic Regression, Bayesian classifier, Gradient Boosting.
TRIM 29	Random Forest
KRT17	Random Forest, Least Angle Regression.
RCN2	Random Forest
SERPINB5	Random Forest, Least Angle Regression.
MIA	Random Forest, Least Angle Regression.
PLEKHB1	Random Forest, Least Angle Regression.
SPDEF	Random Forest
KCNK5	Random Forest

MYC	Random Forest
KRT16	Least Angle Regression
FERMT1	Least Angle Regression.
VTCN1	Least Angle Regression.
DSG2	Least Angle Regression.
KRT6B	Least Angle Regression.
EPHB3	Least Angle Regression.
VGLL1	Least Angle Regression.
SERP1	Least Angle Regression.
FOXC1	Least Angle Regression.
KRT5	Least Angle Regression.

Discussion

Based on the results, ROPN1B and GABRP are the most correlated genes where ROPN1B shows as the most relevant gene to BLBC both in Gradient Boosting and Bayesian models with the average of 96.01% accuracy, and third associated gene in Logistic regression and Decision tree with average of 95.87% accuracy. GABRP also is a robust, relevant gene to BLBC; it is the first important gene in both Logistic regression and Decision tree with an average of 95.87% accuracy, and the 2ed in Gradient Boosting and Bayesian models with average of 96.01% accuracy. However, the Gradient Boosting model shows the highest accuracy of 0.971591 compared to other methods.

We systematically searched the web of science databases, PubMed, and Journals to identify studies which support our results. Table 8 represents the relevant studies for each gene. In the included studies, the expression of ROPN1B, GABRP, JUP, DSC2, TMEM5, PLA2G1B, TRIM 29, RCN2, EPHB3, SERPINB5, MIA, SPDEF, KCNK5, MYC, KRT5, KRT16, KRT6B, KRT17, FERMT1, EPHB3, VGLL1, SFRP1, and

FOXC1 were up-regulated in breast cancer as general while the down-regulated CYB5R1, DLG5, and C19orf73 expressions were associated with BC.

Table 8. Gene's list with relevant studies supports our results.

Gene	Relevant studies
ROPN1B	According to Jai Mehta study[105], ROPN1B was significantly up-regulated in breast cancer patients who relapsed, BC patients who did not survive for more than five years, BC patients who relapsed within five years, and patients with negative Estrogen Receptor tumors. Besides, ROPN1B up-regulated in a sub-group of ER-negative BC with a high incidence of relapse. Another study[106] also reported an up-regulated expression of ROPN1B in $ER^-/HER2^-$ BC tumors. However, It is it is crucial to investigate the role of ROPN1B in BC since little is known about this protein.
GABRP	According to another study[107], about decade ago, GABRP expression was reported to correlate with BLBC. Moreover, GABRP gene not only correlates with BLBC, but also correlated with metastatic dissemination to the brain, showing poorer prognosis. They reported that silencing GABRP in BLBC cells decreased migration, BLBC-associated cytokeratins, and ERK1/2 activation. Furthermore, GABRP expression was up-regulated in $ER^-/HER2^-$ BC tumors based on Bioinformatics

[105] Jai Mehta, 2010.

[106] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y. Identification of key candidate genes, pathways and related prognostic values in ER-negative/HER2-negative breast cancer by bioinformatics analysis. J BUON. (2018) 891-901

[107] Sizemore, Gina M., Steven T. Sizemore, Darcie D. Seachrist, and Ruth A. Keri, 24102–13.

	analysis.[108] Therefore, knocking down GABRP expression may be a new approach for BLBC treatment.[109]
TGM2	Tissue-type transglutaminase 2(TGM2) is a pro-inflammatory protein associated with the resistance of drugs and the metastatic phenotype in BC. TGM2 reported[110] as an essential link in interleukin (IL)-6 mediated cancer cell aggressiveness, it is also an important mediator of distant metastasis. Suppressing TGM2 appears to increase the chemo-sensitivity of cancer cells that were treated with drugs and could be a therapeutic approach.[111]
JUP	an investigation[112] for breast cancer survival and Plakoglobin (JUP) was done by using multivariate and univariate analyses. JUP might be a function of a double-edged sword molecule. Decrease in JUP expression causes an increase of motility of epithelial cells. Therefore, epithelial-mesenchymal transition is prompted and further cancer metastasis. The same study shows that JUP is an oncogene function. High JUP expression causes in clustered tumor cells with high metastatic potential in BC and reduces the probability of patient survival.

[108] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y, 891-901.

[109] Wali, V. B., Patwardhan, G. A., Pelekanou, V., Karn, T., Cao, J., Ocana, A., . . . Pusztai, L. (2019). Identification and Validation of a Novel Biologics Target in Triple Negative Breast Cancer. *Scientific reports, 9*(1), 1-10.

[110] K. Oh, E. KO, H. Kim. "Transglutaminase 2 facilitates the distant hematogenous metastasis of breast cancer by modulating interleukin-6 in cancer cells". Breast cancer res. 2011.

[111] W He, Z Sun, Z Liu, 1413-1418.

[112] L. Lu, H. Zeng, X. Gu and W. Ma." Circulating tumor cell clusters-associated gene *plakoglobin* and breast cancer survival. Breast cancer Research and Treatment. 491-500.

DSC2	The DSC2 protein is a major component of desmosomes, which provides strength and stability to tissues. A study[113] of gene expression microarray analysis was performed by unsupervised hierarchical clustering to identify relevant genes that was expressed differentially between BLBC and non-BLBC. The results show that DSC2 overexpressed in the TNBC subtype. Furthermore, DSC2 is a part of a six-gene signature that predicts metastasis of BC lung.[114] A later study[115] has been reported that DSC2 is associated with BLBC. This protein has been shown to be highly expressed in TNBCs, being able to significantly predict patient survival, and suggesting their role in the aggressiveness seen in these tumors.[116]
TMEM5	TMEM5 significantly up-regulated in BRCA1, it expressed about three times more in BRCA1 mutated cell line (SL) compared to BRCA1 wild-type cell line (BS). [117] [118]
PLEKHB1	PLEKHB1 known also as KPL-1and KP-1 is a human breast cancer cell line based on the malignant effusion of a patient with breast cancer.[119]

[113] Mathe A, Wong-Brown M, Morten B, Forbes JF, and Braye SG. Novel genes associated with lymph node metastasis in triple negative breast cancer. Sci Rep.

[114] Landemaine, T. *et al.* A six-gene signature predicting breast cancer lung metastasis. *Cancer Res*. (2008)

[115] Culhane. C. and Quackenbush, J. Confounding effects in "A six-gene signature predicting breast cancer lung metastasis". *Cancer Res*. (2009).

[116] Hill, J. J., Tremblay, T. L., Fauteux, F., Li, J., 1376-1388

[117] Han, W., Jung, E. M., Cho, J., Lee, J. W., Hwang, K. T., Yang, S. J., . . . Park, I. A.,490-499.

[118] Privat M, Rudewicz J, Sonnier N, and Tamisier C. Antioxydation and Cell Migration Genes Are Identified as Potential Therapeutic Targets in Basal-Like and BRCA1 Muted Breast Cancer Cell Lines. Int J Med Sci. 46-58 (2018).

[119] J Kurebayashi , M Kurosumi and H Sonoo. A new human breast cancer cell line, KPL-1 secretes tumor-associated antigens and grows rapidly in female athymic nude mice. Br J Cancer. (1995) 845-853.

GMDS	GMDS is involved in the process of cellular fucosylation of glycoproteins, which is involved in the functional regulation of adhesion molecules and growth factor receptors, with high levels of fucusylation being reported in various types of cancer.[120] This has been associated with TNBC and EMT, making GMDS a potential player in this process.[121]
CYB5R1	Estrogen-related receptor alpha (ERRα) is overexpressed in different types of tumors, including breast tumors (tripe-negative breast cancer). It is associated with more aggressive tumors, worse outcomes, and increased rate of recurrence.[122] Taken together with another study[123] suggested that CYB5R1 was significantly down-regulated based on microarray analysis of ERRα-silenced HCT116 cells. Along with our results, we suggest that CYB5R1 may correlate with TNBC, Basal-like breast cancer.
DLG5	DLG5 is another gene that correlated with BLBC whose loss of expression resulted in Hippo pathway inhibition through the induction of Scribble mislocalization and down regulating its expression. Also, loss of DLG5 leads to increasing Yes-associated protein(YAP) nuclear localization; in summary, loss of DLG5 expression promoted breast cancer malignancy so

[120] Miyoshi, E., Moriwaki, K., & Nakagawa, T. (2008). Biological function of fucosylation in cancer biology. *J Biochem, 143*(6), 725-729.

[121] Listinsky, J. J., Siegal, G. P., & Listinsky, C. M., 292-322.

[122] Berman AY, Manna S and Schwartz NS. ERRα regulates the growth of triple-negative breast cancer cells via S6K1-dependent mechanism. (2017)

[123] Bernatchez G, Giroux V and Lassalle T ERRα metabolic nuclear receptor controls growth of colon cancer cells. Carcinogenesis 34(10):2253–61. (2013).

	a more effective tumor therapy can be achieved by over expression of DLG5.[124] This study along with previous studies show that DLG5 acts as a tumor suppressor in breast cancer[125], and its expression is up-regulated in BLBC according to our study.
PLA2G1B	The overexpression of PLA2G1B is associated with a high level of Choline. However, a high level of Choline and Phosphocholine have been demonstrated in BC cells. Worth to mention, most of the xenograft models were BLBC.[126]
C19orf73	Huayan et al.[127] examined the integrin β4 expression in breast tumors, and its function in cancer stem cells regulation. As it is known that integrin β4 contributes to BC tumors in terms of invasion, formation, and metastasis. According to her data, β4 expression expressed heterogeneously in BC; but not directly expressed in cancer stem cells but correlated with the population of basal epithelial. However, c19orf73 observed as down-regulated in β4 knockout cells.
NFIX	A study[128] used a DNA methylation MIRA microarray analysis to identify biomarkers for early detection of BC. According to their results,

[124] Liu, J., Li, J., Li, P., Wang, Y., Liang, Z., Jiang, Y., . . . Chen, H. (2017), 42125.

[125] Lu, H., Wang, H., & Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Expert Systems with Applications, 116*, 340-350.

[126] Grinde, M.T., Skrbo, N., Moestue, S.A. et al. Breast Cancer Res (2014) 16: R5. https://doi.org/10.1186/bcr3597

[127] Huayan, S. *et al.* Function of the β4 Integrin in Cancer Stem Cells and Tumor Formation in Breast Cancer: A Masters Thesis. Doi:10.13028/M2588G. (2016)

[128] Lian ZQ, Wang Q and Li WP. Screening of significantly hypermethylated genes in breast cancer using microarray-based methylated-CpG island recovery assay and identification of their expression levels. Int J Oncol. (2012)

	NFIX was one of three genes that was first identified to be hypermethylated in BC.
TRIM 29	A study reported that an overexpression of exogenous microRNA-761 amplified the TNBC cell proliferation, migration, invasion, and colony formation in vivo. MicroRNA-761 represses TRIM29 expression, inducing aggressive phenotypes in TNBC cells. However, the overexpression of TRIM29 reversed the TNBC cell proliferation and invasion.[129] TRIM29 may act as a regulatory factor in cancer tumors. In support of this research, a recent study had been reported that the up-regulated of TRIM29 expression is associated with ER^-/$HER2^-$ BC tumors.[130] Using TRIM29, there needs to be more studies conducted in breast cancer cells because TRIM29 suppresses invasiveness by down-regulating the expression of TWIST1, whereas TRIM29 promotes cell invasion by regulating MMP-9 in lung cancer.[131]
RCN2	RCN2 indicated in many reports an up-regulated in various types of cancer tumors including breast, colorectal, kidney, and liver cancer.[132]
SERPINB5 (MASPIN)	Many studies claim that the loss of SERPINB5 expression is associated with breast cancer. One of the studies[133] reported that the epithelial gene

[129] Guo, Guang-Cheng, Jia-Xiang Wang, Ming-Li Han, Lian-Ping Zhang, and Lin Li., 157–66.

[130] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y, 891-901.

[131] Hatakeyama, S. (2016). Early evidence for the role of TRIM29 in multiple cancer models. In: Taylor & Francis.

[132] Xu, S., Xu, Y., Chen, L. *et al.* RCN1 suppresses ER stress-induced apoptosis via calcium homeostasis and PERK–CHOP signaling. *Oncogenesis* **6.** (2017)

[133] Vecchi, M., Confalonieri, S and Nuciforo, P. *et al.* Breast cancer metastases are molecularly distinct from their primary tumors. *Oncogene* **27,** (2008)

	SERPINB5 significantly inhibited cell motility. Another report[134] suggested that SERPINB5 has involved in determining the metastatic potential of BC cell lines. Furthermore, SERPINB5 expression was reported to correlate with BLBC rather than to be a myoepithelial markers in TNBC. MASPIN may play a substantial role in regulating processes that are associated with the progression and metastatic cascade of TNBC and could present an exclusive and specific target for the diagnosis and therapeutic intervention of TNBC.
MIA	(MIA) is known as a small secreted protein expressed in cartilage; a recent study reported that it is overexpressed in breast cancer.[135]In *situ* expression patterns study[136], MIA expression has been observed at higher levels in breast cancer and reported to have a much bolder expression in malignant epithelial neoplasm. A recent study supports the correlation link between MIA and TNBC by reporting an overexpression of MIA in ER$^-$/HER2$^-$ BC tumors.[137]
SPDEF	Androgen receptor expression overexpressed in approximately 70% of breast cancer. SPDEF is one of the Androgen receptor-related genes, which reported as overexpressed in molecular apocrine tumors.[138] However, PDEF expression restricted to epithelial cells in the breast.[139]

[134] Umekita, Y, Ohi, Y and Souda, M. *et al.* Maspin expression is frequent and correlates with basal markers in triple-negative breast cancer. *Diagn Pathol* **6.** (2011)

[135] Bosserhoff AK, Moser M, Hein R, Landthaler M, Buettner R. (1999). *J Pathol* **187**: 446–454.

[136] Ibid

[137] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y, 891-901.

[138] Lehmann-Che, J., Hamy, A. and Porcher, R. *et al.* Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either HER2 or GCDFP15. *Breast Cancer Res,* **15.** (2013)

[139] Steffan JJ and Koul HK: Prostate derived ETS factor (PDEF): a putative tumor metastasis suppressor. Cancer Lett. (2011) 109-117.

KCNK5	Overexpression of KCNK5 has been observed to have a significant correlation with TNBC; however, it failed to meet the significance criteria to be relevant to BLBC.[140] Conversely, Clarker et al. suggested that up-regulated KCNK5 expression is associated with poor outcomes in BLBC.[141]
MYC	Based on Immunohistochemical analysis,[142] high expression of MYC was reported to associate with the BLBC tumor subtype. Furthermore, other studies of the transformation of epithelial cells show the correlation between MYC and BLBC.[143] [144]
KRT5, KRT16, KRT6B and KRT17	KRT5, KRT16, KRT6B and KRT17 are early detection biomarkers for TNBC tumors known as Basal-like cytokeratins.[145] A recent investigation[146] found that KRT6B, KRT16, KRT17, and KRT81 were up-regulated and correlated with cancer cell proliferation and invasion pathways. Further support of the link between KRT16 and BLBC, a

[140] Dookeran, K. A., Zhang, W., Stayner, L and Argos, M. Associations of two-pore domain potassium channels and triple negative breast cancer subtype in the cancer genome atlas: systematic evaluation of gene expression and methylation. *BMC Res.* (2017)

[141] Clarke C, Madden SF, Doolan P and Aherne ST, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis. (2013)

[142] Chandriani, S. et al. A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. *PLoS ONE* **4**.(2009)

[143] CM Perou CM, SS Jeffrey, M van de Rijn, CA Rees and MB Eisen, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U (1999).9212–9217

[144] CM Perou, T Sorlie, MB Eisen, M van de Rijn and SS Jeffrey, et al. Molecular portraits of human breast tumours. Nature 406. (2000)747–752

[145] BD. Lehmann, JA Bauer and X Chen , et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. (2011) 2750–2767.

[146] Zhang P, Zheng P, and Yang L. Amplication of the CD24 Gene Is an Independent Predictor for Poor Prognosis of Breast Cancer. Front Genet. 2019; 10:560.

	study[147] of circulating tumor cells reported that overexpression of KRT16 in BLBC cell lines was associated with shorter relapse-free survival.
FERMT1	Based on bioinformatics analysis from microarray data, FERMT1 up-regulated in ER-negative/HER2-negative breast cancer tumors.[148] Another study[149] reported that FERMT1 was predictive of BC lung metastases, which in gene expression of 23 metastases of BC tumors were analyzed.
VTCN1	A report[150] suggests that VTCN1 expression was significantly different in BC tumors compared to normal tumors, which may be involved in the progression of BC and metastasis. Also, they suggested that VTCN1 expression could be an early-biomarker for BC.
DSG2	DSG2 expression was reported to present in the invasion and motility of BC cells; also, it may act as a tumor suppressor molecule.[151]
EPHB3	Ephrin B receptors are associated with complex signally pathways in cancer. A study conducted by using microarray data for 3,554 patients had reported that overexpression of EPHB3 was significantly associated with worse survival in BC patients. From the same study findings,

[147] Joosse SA, Hannemann J and Spotter J, et al. Changes in keratin expression during metastatic progression of breast cancer: impact on the detection of circulating tumor cells. Clin Cancer Res. (2012)993-1003

[148] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y, 891-901.

[149] Culhane. C. and Quackenbush, J, 2009.

[150] Tsai SM, Wu SH and Hou MF. The Immune Regulation VTCN1 Gene Polymprphisms and Its Impact in Susceptibility to Breast Cancer. J Clin Lab Anal. (2015)

[151] E Davies. *et al.* The role of desmoglein 2 and E-cadherin in the invasion and motility of human breast cancer cells. *Int. J. Oncol,***11**. (1997) 415–419

	EPHB3 was also reported to have an association with the improvement of relapse-free survival with BLBC.
VGLL1	According to bioinformatics analysis results, Overexpression of VGLL1 associated with ER-negative/HER2-negative breast cancer.[152]
SFRP1	SFRP1 has been suggested to be a potential prognostic marker[153] . Moreover, an increased secretion of this protein has correlated with higher expression in BLBC cell lines. Furthermore, an study showed that SFRP1 strongly correlates with the TNBC subtype and that SFRP1 might be used as a marker classifying patients to positively respond to neoadjuvant chemotherapy.[154]
PADI2	PADI2 expression correlates with breast cancer, where PADI2 expression contributes to migration of abnormal in breast cancer tumor cells.[155]
HOXD13	It is related with positive LNM and tumor size in breast cancer. In addition, low levels of HOXD13 correlates with poorer survival in breast cancer patients.[156]

[152] Shao N, Yuan K, Zhang Y, Yun Cheang T, Li J and Lin Y, 891-901.

[153] Huelsewig, 2014

[154] Bernemann, Christof, Carolin Hülsewig and Christian Ruckert et al, 2014.

[155] Wang, H., Xu, B., Zhang, X., Zheng, Y., Zhao, Y., & Chang, X. (2016). PADI2 gene confers susceptibility to breast cancer and plays tumorigenic role via ACSL4, BINC3 and CA9 signaling. *Cancer cell international, 16*(1), 61.

[156] Zhong, Z.-B., Shan, M., Qian, C., Liu, T., Shi, Q.-Y., Wang, J., . . . Pang, D. (2015). Prognostic significance of HOXD13 expression in human breast cancer. *International journal of clinical and experimental pathology, 8*(9), 11407.

TM4SF1	A high level of TM4SF1 was shown in TNBC tissues, suggesting that TM4SF1 might be a biomarker of TBNC.[157]
FOXC1	A higher level of FOXC1 expression boosts TNBC metastasis by activating the transcription of CXRC4. However, in a zebrafish tumor, either siRNA or AMD3100 in MDA-MB-231 cells can inhibit CXRC4 by down-regulated FOXC1.[158] According to an analysis of microarray data sets for 2,073 breast cancer patients, FOXC1 suggested to be a diagnostic and prognostic biomarker, and could be used as a therapeutic target for the BLBC subtype.[159] FOXC1 is a desirable avenue for further research as a possible therapeutic target in cancer treatment using nanomedicine.

Another study shows that UGT2B28 expression in a breast cancer cell line, suggests its role in androgen and estrogen metabolism.[160] There is currently no study showing a correction between UGT2B28 and the TNBC subtype. According to related studies, UCP1, 2 and 3 act as inductors for autophagy and mitochondrial dysfunction in breast cancer cells, which cause a significant reduction in tumor growth.[161] However, as with

[157] Xing, P., Dong, H., Liu, Q., Zhao, T., Yao, F., Xu, Y., . . . Jin, F. (2017). Upregulation of transmembrane 4 L6 family member 1 predicts poor prognosis in invasive breast cancer. *Medicine, 96*(52), e9476-e9476

[158] Pan, Hongchao, Zhilan Peng, Jiediao Lin and Xiaosha Ren, 3794–3804.

[159] P.S. Ray., *et al.* FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res.*​**70**. (2010) 3870–3876.

[160] Farrar, L. B., Kinyamu, H. K., Flintosh, N., Archer, T. K., & Grant, D. J. (2006). The Regulation of the UDP-glucuronosyltransferase 2B28 gene by glucocorticoids and epidermal growth factor. In: AACR.

[161] Sanchez-Alvarez, R., Martinez-Outschoorn, U. E., Lamb, R., Hulit, J., Howell, A., Gandara, R., . . . Sotgia, F. (2013). Mitochondrial dysfunction in breast cancer cells prevents tumor growth: understanding chemoprevention with metformin. *Cell Cycle, 12*(1), 172-182.

UGT2B28, there are no published results available to support UPC2's correlation with either TNBC or BLBC.

Conclusions

Triple-negative breast cancer (TNBC) constitutes approximately 20%-25% of all breast cancer cases with poor prognoses, with Basal-like breast cancer (BLBC) being a subtype representing 72.8% of TNBC. Classifying BLBC subtypes is of paramount importance for proper diagnosis, with direct clinical implications by dictating the most effective course of treatment.

Although prior research has shown that profiling breast cancers using gene expression data has been useful in investigating and defining prognosis and therapy, little attention has been paid to the molecular characteristics of the basal-like group of breast cancers. Most (if not all) microarray studies of BLBC have been based on small sample size and conducted in isolation from one another in most cases, thus limiting the generalizability of the results. To illustrate the significance of data integration in microarray gene profiling of basal-like breast cancers. in this study, we combined over 24,000 genes of 579 TNBC patients from several TNBC gene expression datasets to identify several important gene signatures in BLBC. A series of different predictive models were built to analyze the data with acceptable accuracy rates. The high dimensionality of the resultant dataset negatively affected the models' performance due to overfitting. To address this issue, several feature selection algorithms were applied to the combined microarray data in order to identify informative genes for building predictive models. Our results show the usefulness of data integration in finer understanding of gene expression in basal-like breast cancers. In addition, a combination of data mining and feature selection techniques allow new genes related to basal-like breast cancers to be identified from many data sources that may be otherwise difficult to detect. In particular, our results showed that the most important genes that correlate with BLBC are ROPN1B and GABRP, SERPINB5, FOXC1, KRT16 and KRT17. Our analysis provided new insights into the pathways in the basal-like group of breast cancers which need to be further investigated in order to develop BLBC specific treatments. The primary focus of different therapeutic approaches for cancer treatment is cancer cells apoptosis. Nanomedicines may be the treatment of choice for all the different types of cancer due to their excellent efficacy in penetration, specific retention and killing of tumor cells. However, the success of nanomedicine is the specific markers and signatures of the cancer cells which can be achieved by analyzing gene datasets.

Reference

1. Agnihotri, N., Kumar, S., & Mehta, K. (2013).Tissue transglutaminase as a central mediator in inflammation-induced progression of breast cancer. *Breast Cancer Research, 15*, 202.
2. Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn Jr, C. E., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer, 116*(14), 3310-3321.
3. Badowska-Kozakiewicz AM and Budzik MP (2016). Immunohisto-chemical characteristics of basal-like breast cancer. Contemp Oncol (Pozn). 20:436–443.
4. Barrett, T., & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol, 411*, 352-369. doi:10.1016/S0076-6879(06)11019-8
5. Bauer KR, Brown M, Cress RD, et al (2007). Descriptive analysis of estrogen receptor (ER)-negative, pro-gesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the California Cancer Registry. Cancer 109:1721-1728.
6. Becker-Santos, D. D., Lonergan, K. M., Gronostajski, R. M., & Lam, W. L. (2017). Nuclear factor I/B: a master regulator of cell differentiation with paradoxical roles in cancer. *EBioMedicine, 22*, 2-9.
7. Berman AY, Manna S, Schwartz NS, Katz YE, Sun Y, Behrmann CA, Yu JJ, Plas DR, Alayev A, Holz MK. (2017). ERRα regulates the growth of triple-negative breast cancer cells via S6K1-dependent mechanism. Signal Transduct Target Ther 2: e17035.
8. Bernatchez G, Giroux V, Lassalle T, Carpentier AC, Rivard N, Carrier JC. (2013). ERRα metabolic nuclear receptor controls growth of colon cancer cells. Carcinogenesis 34(10):2253–61.
9. Bernemann, Christof, Carolin Hülsewig, Christian Ruckert, Sarah Schäfer, LenaBlümel, Georg Hempel, Martin Götte, et al (2014) "Influence of Secreted Frizzled Receptor Protein 1 (SFRP1) on Neoadjuvant Chemotherapy in Triple Negative Breast Cancer Does Not Rely on WNT Signaling." *Molecular Cancer* 13, no. 1: 174.
10. Bosserhoff AK, Moser M, Hein R, Landthaler M, Buettner R. (1999). *In situ* expression patterns of melanoma-inhibiting activity (MIA) in melanomas and breast cancers. *J Pathol*, 187: 446– 54.
11. Boyle P (2012). Triple-negative breast cancer: epidemiological considerations and recommendations. *Ann Oncol*, 23, 7-12.
12. Carey LA, Perou CM, Livasy CA, et al (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA*, 295, 2492-502.
13. Chandriani, S. et al. (2009). A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. *PLoS ONE* **4**, e6693.
14. Cheang MC, Voduc D, Bajdik C, *et al* (2008). Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin Cancer Res*, 14:1368–1376
15. Chen, A. H., & Yang, C. (2012). The improvement of breast cancer prognosis accuracy from integrated gene expression and clinical data. *Expert Systems with Applications, 39*(5), 4785-4795.

16. Chen, H., Wu, J., Zhang, Z., Tang, Y., Li, X., Liu, S., . . . Li, X. (2018). Association Between BRCA Status and Triple-Negative Breast Cancer: A Meta-Analysis. *Front Pharmacol, 9*, 909. doi:10.3389/fphar.2018.00909
17. Cheng, Zhongle, Wei Wei, Zhengshen Wu, Jing Wang, Xiaojuan Ding, Youjing Sheng, Yinli Han, and Qiang Wu (2019). "ARPC2 Promotes Breast Cancer Proliferation and Metastasis." *Oncology Reports*,41, 3189-3200.
18. Choi J, Jung WH, Koo JS (2012). Clinicopathologic features of molecular subtypes of triple negative breast cancer based on immunohistochemical markers. *Histol Histopathol*, 27, 1481-93.
19. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J, et al. (2013) Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis, 34, 2300–8.
20. Culhane, A. C. & Quackenbush, J. (2009). Confounding effects in "A six-gene signature predicting breast cancer lung metastasis". *Cancer Res* **69**, 7480–5.
21. Czerwińska, Patrycja, Parantu K. Shah, Katarzyna Tomczak, Marta Klimczak, Sylwia Mazurek, Barbara Sozańska, Przemysław Biecek, et al (2016). "TRIM28 Multi-Domain Protein Regulates Cancer Stem Cell Population in Breast Tumor Development." *Oncotarget,* 8, 863-882.
22. Davies, E. *et al.* (1997). The role of desmoglein 2 and E-cadherin in the invasion and motility of human breast cancer cells. *Int. J. Oncol.***11**, 415–419.
23. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine, 34*, 113-127
24. Dookeran, K. A., Zhang, W., Stayner, L., and Argos, M. (2017). Associations of two-pore domain potassium channels and triple negative breast cancer subtype in the cancer genome atlas: systematic evaluation of gene expression and methylation. *BMC Res. Notes* 10:475. doi: 10.1186/s13104-017-2777-2774
25. Fan, Chunni, Ning Liu, Dan Zheng, Jianshi Du, and Keren Wang (2019). "MicroRNA-206 Inhibits Metastasis of Triple-Negative Breast Cancer by Targeting Transmembrane 4 L6 Family Member 1." *Cancer Management and Research,* 11, 6755–64.
26. Farrar, L. B., Kinyamu, H. K., Flintosh, N., Archer, T. K., & Grant, D. J. (2006). The Regulation of the UDP-glucuronosyltransferase 2B28 gene by glucocorticoids and epidermal growth factor. In: AACR.
27. Friedenson, Bernard (2007). "The BRCA1/2 Pathway Prevents Hematologic Cancers in Addition to Breast and Ovarian Cancers." *BMC Cancer* 7, no. 1. doi:10.1186/1471-2407-7-152.
28. G. Tzanis, C. Berberidis, and I. Vlahavas, "Biological data mining," *Scientific Programming*, no. 16, 2015. [Online]. Available: https: //www.researchgate.net/publication/220060935 Biological Data Mining.
29. Grinde, M.T., Skrbo, N., Moestue, S.A. et al. Breast Cancer Res (2014) 16: R5. https://doi.org/10.1186/bcr3597
30. Guo, Guang-Cheng, Jia-Xiang Wang, Ming-Li Han, Lian-Ping Zhang, and Lin Li. (2017). "microRNA-761 Induces Aggressive Phenotypes in Triple-Negative Breast Cancer Cells by Repressing TRIM29 Expression." *Cellular Oncology,* 40. 157–66.
31. H.M. Zolbanin et al. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. Decision Support Systems 74, 150–161.

32. Hájek P, (2011). Municipal credit rating modelling by neural networks, Decision Support Systems 51, 108–118.
33. Han, J., (2002). "How can data mining help bio-data analysis"? In: Zaki, M.J., Wang, J.T.L. and Toivonen, H.T.T. (Eds). Proceedings of the 2nd ACM SIGKDD Workshop on data mining in bioinformatics, Vol. 1-2.
34. Han, W., Jung, E. M., Cho, J., Lee, J. W., Hwang, K. T., Yang, S. J., . . . Park, I. A. (2008). DNA copy number alterations and expression of relevant genes in triple-negative breast cancer. *Genes, Chromosomes and Cancer, 47*(6), 490-499.
35. He, W., Sun, Z., & Liu, Z. (2015). Silencing of TGM2 reverses epithelial to mesenchymal transition and modulates the chemosensitivity of breast cancer to docetaxel. *Exp Ther Med, **10***, 1413-1418. doi:10.3892/etm.2015.2679
36. Hill, J. J., Tremblay, T. L., Fauteux, F., Li, J., Wang, E., Aguilar-Mahecha, A., . . . O'Connor-McCourt, M. (2015). Glycoproteomic comparison of clinical triple-negative and luminal breast tumors. *J Proteome Res, 14*(3), 1376-1388. doi:10.1021/pr500987r
37. Holder, Ashley M., Ana M. Gonzalez-Angulo and Huiqin Chen. (2012). "High Stearoyl-CoA Desaturase 1 Expression Is Associated with Shorter Survival in Breast Cancer Patients." *Breast Cancer Research and Treatment* 137, no. 1, 319–27.
38. Hosokawa, Yuko, Noritaka Masaki, Shiro Takei and Makoto Horikawa. (2017). "Recurrent Triple-Negative Breast Cancer (TNBC) Tissues Contain a Higher Amount of Phosphatidylcholine (32:1) than Non-Recurrent TNBC Tissues." Edited by Irina U. Agoulnik. *PLOS ONE* 12, no. 8, e0183724.
39. Huayan, S. *et al.* (2016). Function of the β4 Integrin in Cancer Stem Cells and Tumor Formation in Breast Cancer: A Masters Thesis. Doi:10.13028/M2588G.
40. Huelsewig, Carolin, Christof Bernemann and Christian Ruckert. (2014). "Abstract 920 Secreted Frizzled Related Protein 1 (SFRP1) as Potential Regulator of Chemotherapy Response for Patients with Triple Negative Breast Cancer (TNBC)." In Clinical Research (Excluding Clinical Trials). *American Association for Cancer Research*, 27, 464-477.
41. J. P. Mehta, "Gene expression analysis in breast cancer," Ph.D. dissertation, Dublin City University, Ireland, 2010.
42. Joosse SA, Hannemann J and Spotter J et al. (2012). Changes in keratin expression during metastatic progression of breast cancer: impact on the detection of circulating tumor cells. Clin Cancer Res, 993-1003.
43. Kim, W., Kim, K. S., Lee, J. E., Noh, D.-Y., Kim, S.-W., Jung, Y. S., . . . Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer, 15*(2), 230-238.
44. Kurebayashi J, Kurosumi M, Sonoo H. (1995). A new human breast cancer cell line, KPL-1 secretes tumor-associated antigens and grows rapidly in female athymic nude mice. *Br J Cancer*, 71: 845-853.
45. Kuroda, Naoto, Masahiko Ohara, Kaori Inoue and Keiko Mizuno. (2009). "The Majority of Triple-Negative Breast Cancer May Correspond to Basal-like Carcinoma, but Triple-Negative Breast Cancer Is Not Identical to Basal-like Carcinoma." *Medical Molecular Morphology* 42,128–31.
46. Landemaine, T. *et al.* (2008). A six-gene signature predicting breast cancer lung metastasis. *Cancer Res* **68**, 6092–9.

47. Lehmann BD, Bauer JA and Chen X, et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest*, 2750–2767.
48. Lehmann-Che, J., Hamy, A., Porcher, R. *et al.* (2013). Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either HER2 or GCDFP15. *Breast Cancer Res,* 15**,** R37 doi:10.1186/bcr3421
49. Lian ZQ, Wang Q, Li WP, Zhang AQ, Wu L. (2012). Screening of significantly hypermethylated genes in breast cancer using microarray-based methylated-CpG island recovery assay and identification of their expression levels. *Int J Oncol*, 41(2):629–38.
50. Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M. (2008). Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol*, 26:1275-1281
51. Listinsky, J. J., Siegal, G. P., & Listinsky, C. M. (2011). The emerging importance of alpha-L-fucose in human breast cancer: a review. *Am J Transl Res, 3*(4), 292-322.
52. Liu Y, Xin T, Jiang QY, et al (2013). CD147, MMP9 expression and clinical significance of basal-like breast cancer. *Med Oncol*, 33:366.
53. Liu, J., Li, J., Li, P., Wang, Y., Liang, Z., Jiang, Y., . . . Chen, H. (2017). Loss of DLG5 promotes breast cancer malignancy by inhibiting the Hippo signaling pathway. *Scientific reports, 7*, 42125.
54. Liu, J., Li, J., Ren, Y., & Liu, P. (2014). DLG5 in cell polarity maintenance and cancer development. *International journal of biological sciences, 10*(5), 543.
55. Lu L, Zeng H, Gu X and Ma W. (2015).” Circulating tumor cell clusters-associated gene *plakoglobin* and breast cancer survival. *Breast cancer Research and Treatment*. 491-500.
56. Lu, H., Wang, H., & Yoon, S. W. (2019). A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. *Expert Systems with Applications, 116*, 340-350.
57. M. Palwal, U.Kumar. (2009). Neural networks and statistical techniques: A review of applications Expert Systems with Application. 36(1),2-17
58. Macdonald F, CHJ Ford, Casson AG.(2004). Breast cancer. In ‘Molecular Biology of Cancer’, Eds Macdonald F, CHJ Ford, Casson AG. *BIOS Scientific Publishers, London and New York*, 139-63.
59. Mathe A, Wong-Brown M, Morten B, Forbes JF, Braye SG, Avery-Kiejda KA, Scott RJ. (2015). Novel genes associated with lymph node metastasis in triple negative breast cancer. Sci Rep, 5:15832.
60. Miyoshi, E., Moriwaki, K., & Nakagawa, T. (2008). Biological function of fucosylation in cancer biology. *J Biochem, 143*(6), 725-729. doi:10.1093/jb/mvn011
61. Miyoshi, E., Moriwaki, K., & Nakagawa, T. (2008). Biological function of fucosylation in cancer biology. *J Biochem, 143*(6), 725-729. doi:10.1093/jb/mvn011
62. Nathanson KN, Wooster, R, Weber, BL. (2001).Breast cancer genetics: what we know and what we need. *Nat. Med.,* 7 (2001),552-556
63. Oh, K., Ko, E., Kim, H.S. *et al.* (2011). Transglutaminase 2 facilitates the distant hematogenous metastasis of breast cancer by modulating interleukin-6 in cancer cells. *Breast Cancer Res* **13,** R96. doi:10.1186/bcr3034
64. Palmieri, V., Bozzi, M., Signorino, G., Papi, M., De Spirito, M., Brancaccio, A., . . . Sciandra, F. (2017). alpha-Dystroglycan hypoglycosylation affects cell migration by

influencing beta-dystroglycan membrane clustering and filopodia length: A multiscale confocal microscopy analysis. *Biochim Biophys Acta Mol Basis Dis, 1863*(9), 2182-2191. doi:10.1016/j.bbadis.2017.05.025

65. Pan, Hongchao, Zhilan Peng and Jiediao Lin. (2018). "Forkhead Box C1 Boosts Triple-Negative Breast Cancer Metastasis through Activating the Transcription of Chemokine Receptor-4." *Cancer Science* 109, 3794–3804.
66. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, et al. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A 96: 9212–9217
67. Perou CM. (2011). Molecular stratification of triple-negative breast cancers. *Oncologist* 16 (suppl. 1). 61–70
68. Perou, C., Sørlie, T., Eisen, M. *et al.* Molecular portraits of human breast tumours. *Nature* 406**,** 747–752 (2000) doi:10.1038/35021093
69. Peshkin, B. N., Alabek, M. L., & Isaacs, C. (2010). BRCA1/2 mutations and triple negative breast cancers. *Breast Dis, 32*(1-2), 25-33. doi:10.3233/BD-2010-0306
70. Piatetsky-Shapiro, G. and Tamayo, P. (2003). Microarray Data Mining: Facing the Challenges. SIGKDD Explorations, 5(2), 1-5.
71. Prat A, Adamo B, Cheang MC. (2013). Molecular Characterization of basal-like and non-basal-like triple negative breast cancer. *Oncologist*, 18: 123-133
72. Privat M, Rudewicz J, Sonnier N, Tamisier C, Ponelle-Charchuat F, Bignon YJ. (2018). Antioxydation and Cell Migration Genes Are Identified as Potential Therapeutic Targets in Basal-Like and BRCA1 Mutated Breast Cancer Cell Lines. Int J Med Sci. 1;15(1):46-58.
73. Privat, M., Rudewicz, J., Sonnier, N., Tamisier, C., Ponelle-Chachuat, F., & Bignon, Y. J. (2018). Antioxydation And Cell Migration Genes Are Identified as Potential Therapeutic Targets in Basal-Like and BRCA1 Mutated Breast Cancer Cell Lines. *Int J Med Sci, 15*(1), 46-58. doi:10.7150/ijms.20508
74. Ray, P. S. *et al.* (2010). FOXC1 is a potential prognostic biomarker with functional significance in basal-like breast cancer. *Cancer Res.***70**, 3870–3876.
75. Rody et al (2011). A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Research 13: R97.
76. Sanchez-Alvarez, R., Martinez-Outschoorn, U. E., Lamb, R., Hulit, J., Howell, A., Gandara, R., . . . Sotgia, F. (2013). Mitochondrial dysfunction in breast cancer cells prevents tumor growth: understanding chemoprevention with metformin. *Cell Cycle, 12*(1), 172-182.
77. SAS Institute Inc. *SAS Enterprise MinerTM High-Performance Data Mining Node Preference for SAS 9.3.* Cary, NC Institute INC (2011).
78. Scott, K. F., Sajinovic, M., Hein, J., Nixdorf, S., Galettis, P., Liauw, W., . . . Russell, P. J. (2010). Emerging roles for phospholipase A2 enzymes in cancer. *Biochimie, 92*(6), 601-610. doi:10.1016/j.biochi.2010.03.019
79. Scott, K. F., Sajinovic, M., Hein, J., Nixdorf, S., Galettis, P., Liauw, W., . . . Russell, P. J. (2010). Emerging roles for phospholipase A2 enzymes in cancer. *Biochimie, 92*(6), 601-610. doi:10.1016/j.biochi.2010.03.019
80. Shao N, Yuan K, Yun Cheang T, Li J, Lin Y. (2018). Identification of key candidate genes, pathways and related prognostic values in ER-negative/HER2-negative breast cancer by bioinformatics analysis. J BUON. 2018;2 23:891-901

81. Shao, M.-M., Chan, S. K., Yu, A. M. C., Lam, C. C. F., Tsang, J. Y. S., Lui, P. C. W., … Tse, G. M. (2012). Keratin expression in breast cancers. *Virchows Archiv*, 461(3), 313–322.
82. Sizemore, Gina M., Steven Tand Sizemore, Darcie D (2014). "GABA(A) Receptor Pi (GABRP) Stimulates Basal-like Breast Cancer Cell Migration through Activation of Extracellular-Regulated Kinase 1/2 (ERK1/2)." *Journal of Biological Chemistry* 289, no. 35, 24102–13.
83. Sorlie, T., Perou, C. M., Tibshirani, R.(2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 98 (19),10869-10874.
84. Steffan JJ, Koul HK. (2011). Prostate derived ETS factor (PDEF): a putative tumor metastasis suppressor. Cancer Lett, 310: 109-117. 10.1016/j.canlet.2011.06.011.
85. Tan DS, Marchio C, Jones RL, et al (2008). Triple negative breast cancer: molecular profiling and prognostic impact in adjuvant anthracycline-treated patients. *Breast Cancer Res Treat*, 111, 27-44
86. Tsai SM, Wu SH, Hou MF, Yang HH, Tsai LY. (2015). The Immune Regulation VTCN1 Gene Polymprphisms and Its Impact in Susceptibility to Breast Cancer. *J Clin Lab Anal*, 29(5):412-8. Doi: 10.1002/jcla.21788
87. Umekita, Y., Ohi, Y., Souda, M. *et al.* (2011). Maspin expression is frequent and correlates with basal markers in triple-negative breast cancer. *Diagn Pathol* **6,** 36 doi:10.1186/1746-1596-6-36
88. Vecchi, M., Confalonieri, S., Nuciforo, P. *et al.* Breast cancer metastases are molecularly distinct from their primary tumors. *Oncogene* **27,** 2148–2158 (2008) doi: 10.1038/sj.onc.1210858
89. Verma, Mukesh, and Upender Manne.(2006). "Genetic and Epigenetic Biomarkers in Cancer Diagnosis and Identifying High Risk Populations." *Critical Reviews in Oncology/Hematology* 60, no. 1, 9–18.
90. Wali, V. B., Patwardhan, G. A., Pelekanou, V., Karn, T., Cao, J., Ocana, A., . . . Pusztai, L. (2019). Identification and Validation of a Novel Biologics Target in Triple Negative Breast Cancer. *Scientific reports, 9*(1), 1-10.
91. Wang, H., Xu, B., Zhang, X., Zheng, Y., Zhao, Y., & Chang, X. (2016). PADI2 gene confers susceptibility to breast cancer and plays tumorigenic role via ACSL4, BINC3 and CA9 signaling. *Cancer cell international, 16*(1), 61.
92. Woischke, C., Blaj, C., Schmidt, E. M., Lamprecht, S., Engel, J., Hermeking, H., . . . Horst, D. (2016) . CYB5R1 links epithelial-mesenchymal transition and poor prognosis in colorectal cancer. *Oncotarget, 7*, 31350-31360. doi:10.18632/oncotarget.8912
93. Xing, P., Dong, H., Liu, Q., Zhao, T., Yao, F., Xu, Y., . . . Jin, F. (2017). Upregulation of transmembrane 4 L6 family member 1 predicts poor prognosis in invasive breast cancer. *Medicine, 96*(52), e9476-e9476.
94. Xu, S., Xu, Y., Chen, L. *et al.* (2017). RCN1 suppresses ER stress-induced apoptosis via calcium homeostasis and PERK–CHOP signaling. *Oncogenesis* **6,** e304 doi:10.1038/oncsis.2017.6
95. Yadav, Budhi S. (2015). "Biomarkers in Triple Negative Breast Cancer: A Review." *World Journal of Clinical Oncology* 6, 252.

96. Yamashita, S., Ogawa, M., Sakamoto, K., Abe, T., Arakawa, H., & Yamashita, J. (1994). Elevation of serum group II phospholipase A2 levels in patients with advanced cancer. *Clin Chim Acta, 228*(2), 91-99. doi:10.1016/0009-8981(94)90280-1
97. Yang, L., Wu, X., Wang, Y., Zhang, K., Wu, J., Yuan, Y. C., . . . Yen, Y. (2011). FZD7 has a critical role in cell proliferation in triple negative breast cancer. *Oncogene, 30*(43), 4437-4446. doi:10.1038/onc.2011.145
98. Zhang P., Zheng P., Yang L et al. (2019). Amplication of the CD24 Gene Is an Independent Predictor for Poor Prognosis of Breast Cancer. Front Genet, 10:560. Doi: 10.3389/fgene.2019.00560.
99. Zhang, Shizhen, Zhen Wang, Weiwei Liu, Rui Lei, Jinlan Shan, Ling Li, and Xiaochen Wang. (2017). “Distinct Prognostic Values of S100 mRNA Expression in Breast Cancer.” *Scientific Reports* 7, no. 1, 39786.
100. Zhong, Z.-B., Shan, M., Qian, C., Liu, T., Shi, Q.-Y., Wang, J., . . . Pang, D. (2015). Prognostic significance of HOXD13 expression in human breast cancer. *International journal of clinical and experimental pathology, 8*(9), 11407.