

Lift the Veil of Breast Cancers Using Four or Fewer Critical Genes

Zhengjun Zhang

University of Wisconsin

Abstract

Tackling breast cancer problems is like mastering a puzzle, and the mystery is not yet solved. Reported key genes in the literature could not be confirmed whether they are vital to breast cancer formations due to lack of convincing accuracy, although they may be biologically directly related to breast cancer based on present biological knowledge. It is hoped vital genes can be identified with the highest possible accuracy, e.g., 100% accuracy and convincing causal patterns beyond what has been known in breast cancer. One hope is that finding gene-gene interaction signatures and functional effects may solve the puzzle. This research uses a recently developed competing linear factor analysis method in differentially expressed gene detection to advance the study of breast cancer formation to its deepest root level as deep as possible. Surprisingly, three genes are detected to be differentially expressed in TNBC, and non-TNBC (Her2, Luminal A, Luminal B) samples with 100% sensitivity and 100% specificity in one study of triple-negative breast cancers (TNBC, with 54675 genes and 265 samples). These three genes show a clear signature pattern of how TNBC patients can be grouped. For another TNBC study (with 54673 genes and 66 samples), four genes bring the same accuracy of 100% sensitivity and 100% specificity. Four genes are found to have the same accuracy of 100% sensitivity and 100% specificity in one breast cancer study (with 54675 genes and 121 samples), and the same four genes bring an accuracy of 100% sensitivity and 96.5% specificity in the fourth breast cancer study (with 60483 genes and 1217 samples.) These results show the four-gene-based classifiers are robust and accurate. The detected genes naturally classify patients into subtypes, e.g., seven subtypes. These findings demonstrate the clearest gene-gene interaction patterns and functional effects with the smallest numbers of genes and the highest accuracy compared with findings reported in the literature. The four genes are considered to be essential for breast cancer studies and practice. They can provide focused, targeted researches and precision medicine for each subtype of breast cancer. New breast cancer disease types may be detected using the classified subtypes, and hence new effective therapies can be developed.

Keywords: Direct and indirect effects, breast cancer detection, gene-gene interaction, functional effects, joint risk competing.

Introduction

Breast cancer has been an unconquered plague for centuries. It has had the highest death rate among all cancers women have had for many years. It has caused enormous economic losses and costs. To save lives and protect women from breast cancers, enormous research efforts and money have been investigated. Although there have been some considerable signs of progress in breast cancer diagnoses and therapies, many women still suffer from being diagnosed with breast cancer and lost their lives every year. No apparent clues or research results show the most critical genetic causality in breast cancer formation. The most hopeful direction, finding critical genes, or primary differentially expressed genes related to breast cancer formation, has been drawing much attention

in breast cancer studies. The most recent editorial summary by Narod (2021) states "Results of two large case-control studies that analyzed the associations between a number of putative cancer susceptibility genes and breast cancer risk are now reported in the Journal. The study by Dorling et al. (2021, Breast Cancer Association Consortium) included 34 genes and 113,000 women from 25 countries, and the study by Hu et al. (2021) included 28 genes and 64,000 women from the United States. Variants in 8 genes — BRCA1, BRCA2, PALB2, BARD1, RAD51C, RAD51D, ATM, and CHEK2 — had a significant association with breast cancer risk in both studies."

Differential expression analysis between tumor and non-tumor cells helps breast cancer prognosis prediction at a relatively early stage, identifying some clear patterns from patients to patients, recommending different precision therapies according to breast cancer subtypes. Efforts have been made in identifying genes associated with breast cancer symptoms. For example, in a systems biology comprehensive analysis on breast cancer to identify key gene modules and genes associated with TNM-based clinical stages (Amjad et al. 2020), the authors have identified various numbers of genes that can be key genes related to breast cancers at different cancer stages. Malvia et al. (2019) studied gene expression profiles of breast cancers in Indian women, obtained 2413 differentially expressed genes, and demonstrated the existence of molecular subtypes in Indian women. Lv et al. (2019) aimed to explore some novel genes and pathways related to TNBC prognosis through bioinformatics methods as well as potential initiation and progression mechanisms. 755 differentially expressed overlapping mRNAs were detected between TNBC/non-TNBC samples and normal tissue. The authors found eight hub genes associated with the cell cycle pathway highly expressed in TNBC. Additionally, a novel six-gene (TMEM252, PRB2, SMC01, IVL, SMR3B, and COL9A3) signature from the 755 differentially expressed mRNAs were constructed and significantly associated with prognosis as an independent prognostic signature. Zhong et al. (2020) conducted a robust rank aggregation (RRA) analysis based on genome-wide gene expression datasets involving TNBC patients from the Gene Expression Omnibus (GEO) database to identify key genes associated with TNBC. A total of 194 highly ranked differentially expressed genes (DEGs) were identified in TNBC vs. non-TNBC. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analysis was utilized to explore the identified genes' biological functions. The authors also found that some genes are positively correlated to the life expectancy ($P < 0.05$) of TNBC patients. Lin et al. (2020) identified potential key genes for HER-2 positive breast cancer based on bioinformatics analysis. A total of 54 up-regulated DEGs and 269 downregulated DEGs were identified. Among them, 10 hub genes including CCNB1, RAC1, TOP2A, KIF20A, RRM2, ASPM, NUSAP1, BIRC5, BUB1B, and CEP55 demonstrated by connectivity degree in the PPI network were screened out. Chen et al. (2018) systematically searched the electronic databases of MEDLINE (PubMed), Embase, and Cochrane Library to identify relevant publications from April, 1959 to November, 2017. identified 16 qualified studies from 527 publications with 46,870 breast cancer patients including 868 BRCA1 mutations carriers, 739 BRCA2 mutations carriers, and 45,263 non-carriers. The results showed that breast cancer patients with BRCA1Mut carriers were more likely to have TNBC than those of BRCA2Mut carriers (OR: 3.292; 95% CI: 2.773–3.909) or non-carriers (OR: 8.889; 95% CI: 6.925–11.410). Deng et al. (2019) identified potential crucial genes and key pathways in breast cancer using bioinformatic analysis. 203 up-regulated and 118 down-regulated DEGs were identified. Six hub genes were selected and validated in clinical sample for further analysis due to the high degree of connectivity, including CDK1, CCNA2, TOP2A, CCNB1, KIF11, and MELK. They were all correlated to worse overall survival (OS) in breast cancer. Zhu et al. (2020) identified some key genes and pathways associated with irradiation in breast cancer tissue and breast cancer cell lines. A total of 82 DEGs (74 up-regulated and 8 downregulated genes)

were identified. Two characteristic subnetworks and 3 hub genes (FOS, CCL2, and CXCL12) were strongly distinguished in PPI network. Dong et al. (2018) aimed to identify the key pathways and genes and find the potential initiation and progression mechanism of TNBC. 56 up-regulated and 151 downregulated genes were listed and the gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analysis was performed. The authors found that SOX8, AR, C9orf152, NRK and RAB30, and other key genes and pathways might be promising targets for the TNBC treatment. Lu et al. (2019) identified five hub genes (PHLPP1, UBC, ACACB, TGFB1, and ACTB) associated with HER2+BC with brain metastasis. The GSEA analysis revealed that the ribosomal pathway seems to play a very important role in the pathogenesis of HER2+BC with brain metastasis. Such studies are many. This paper won't be able to review all of them. Readers can search and find interesting topics online.

The reported genes in the published work point out some promising directions in breast cancer research and treatments. But it is not clear whether or not they are fundamental causes or direct causes of breast cancer. The problem is mainly due to the following three main limitations. 1) The number of human genes is ultra-large compared to the number of patients in affordable study designs. Identifying a few key (single digit) genes that are uniformly optimal across different trials, different study purposes, different measurement methods, and different cohorts is rather challenging. From the aforementioned research outcomes, we can see there are many different genes are identified. As a result, it's impossible to see which one is the most important one, which can be a driver of breast cancer disease. 2) The inefficient detecting power of existing analysis methods due to restricted model assumptions cannot deal with heterogeneous populations (different breast cancer subtypes). As a result, the sensitivity and specificity of many published gene classifiers are not satisfactory. 3) It isn't easy to extract informative messages from existing models and analysis methods. Also, many gene-related classifiers are not interpretable as gene-gene inter-relationships and functional effects are hardly expressed. As a result, scientific research progress in breast cancer studies is still limited. Much literature attention has been focused on individual genes and their expression levels, i.e., not gene-gene interactions, genes-subtypes (of breast cancers) interactions, and functional effects. As a result, the fundamental genetic causes of breast cancer formations can be masked by those suboptimal focuses, and the researches can still be in a primitive state. Many unknown factors exist. They can be essential to conquer the breast cancer plague, and therefore there is an urgent need for identifying critical DEGs with the highest possible sensitivity and specificity for breast cancer detection.

This work aims to lift the veil of breast cancers by discovering the joint functional effects of four or fewer critical DEGs that show the highest detecting power of breast cancer in four gene expression RNA-seq datasets. According to our analysis, these four genes and their functional effects describe breast cancers' overall features at the genomic level, with the highest possible sensitivity (up to 100%) and specificity (up to 100%) for breast cancer detection. In addition, they are invariance preserving with the same group of patients but measured in different scales, and they are robust from one trial to another trial.

Statistical Methodology

The most recently developed max-linear competing factor models (Cui and Zhang, 2018), max-linear regression models (Cui et al. 2020), and max-linear logistic models (Xu 2019, Zhang 2021) have proven to be powerful models and analysis approaches to study heteroscedastic populations and competing risks and resources. The theoretical foundations of these models have been established in

Cui and Zhang (2018), Cui et al. (2020), Malinowski et al. (2018), Xu (2019), and Zhang (2020, 2021). The difference between the max-linear competing models and the classical statistical models is that the original linear combination of predictors is replaced by the maximum of a set of linear combinations of predictors, called competing factors or competing-risk factors. The max-linear competing factor models are different from existing popular classification models such as random forest, support vector machine, group lasso-based machine learning methods, and deep learning methods. However, the max-linear competing factor models are interpretable and outperform existing methods (Cui et al. 2020). This study implements the max-linear logistic regression model to build a competing factor breast cancer classifier. For completeness, the model is stated as follows.

Suppose there are $i = 1, \dots, n$ patients with breast cancer status label $Y_i = 1$ for cancer and $Y_i = 0$ for cancer-free, and Y_i is related to G groups of genes by

$$\Phi_{ij} = (X_{i,j_1}, X_{i,j_2}, \dots, X_{i,j_{g_j}}), j = 1, \dots, G, g_j \geq 0 \quad (1)$$

where i is the i th individual in the sample, g_j is the number of genes in j th group. The competing (risk) factor classifier for the i th outcome variable is defined as

$$\log\left(\frac{p_i}{1-p_i}\right) = \max(\beta_{01} + \Phi_{i1}\beta_1, \beta_{02} + \Phi_{i2}\beta_2, \dots, \beta_{0G} + \Phi_{iG}\beta_G) \quad (2)$$

where β_{0j} 's are intercepts, Φ_{ij} is a $1 \times g_j$ observed vector, β_j is a $g_j \times 1$ coefficient vector which characterizes the contribution of each predictor to the outcome variable \mathbf{Y} in the j th group to the risk, and $\beta_{0j} + \Phi_{ij}\beta_j$ is called the j th competing risk factor, i.e., j th signature. The unknown parameters are estimated from

$$(\hat{\beta}, \hat{S}) = \arg \min_{\beta, S_j \subset S, j=1,2,\dots,G} \sum_{i=1}^n [I(p_i \leq 0.5)I(Y_i = 1) + I(p_i > 0.5)I(Y_i = 0)]$$

where 0.5 is a probability threshold value that is commonly used in machine learning classifiers, $I(\cdot)$ is an indicate function, p_i is defined in the equation (2). $S = \{1, 2, \dots, 54675\}$ is the index set of all genes, $S_1 = \{1_1, 1_2, \dots, 1_{g_1}\}$, $S_2 = \{2_1, \dots, 2_{g_2}\}$, \dots , $S_G = \{G_1, \dots, G_{g_G}\}$ are index sets

corresponding to (1), and $S = \{1_1, 1_2, \dots, 1_{g_1}; 2_1, \dots, 2_{g_2}; \dots; G_1, \dots, G_{g_G}\}$ is the final gene set selected in the final classifiers.

The goal is to identify the clearest patterns of gene-gene interactions and functional effects related to breast cancer samples and non-tumor samples. We start with three competing factors in max-linear logistic regression models, with each factor having only three genes randomly drawing from 54675, 54673, or 60483 genes. Then, a Monte Carlo method with extensive computation is used to find the final model with the best performance of sensitivity and specificity and the smallest number of genes. Finally, the complete computing description is listed in Zhang (2021) in which five Covid-19 critical genes and seven subtypes were identified.

Data Descriptions

There are four datasets used in this study. The first dataset is triple-negative breast cancer (TNBC, North American cohort) study conducted by Burstein et al. (2015) and den Hollander (2016) with 54675 genes, 198 TNBC tumor samples, and 67 not TNBC (Her2, Luminal A, Luminal B) samples. The data link and descriptions are <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76275>. The platforms are GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. The expression values are $\log_2(\text{RMA signal})$. The second dataset is a European cohort with 55 TNBC

samples and 11 normal breast tissue samples, Maire et al. (2013), Maire et al. (2013), Maubant et al. (2015). The data link and description are <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65194>. The platforms are GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. The number of genes is 54673. The expression values are $\log_2(\text{GCRMA signal from Affy cdf})$. The third dataset is gene expression profiling of 104 breast cancer and 17 normal breast biopsies by Clark et al. (2013). It is from a European cohort. The data link and descriptions are <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42568>. The platforms are GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. The expression values are $\log_2(\text{GC-RMA signal intensity})$. The fourth dataset is GDC TCGA Breast Cancer cohort by Genomic Data Commons. The dataset contains 60,484 identifiers (genes) and 1217 (1104 tumors and 113 tumor free) samples. Data from the same sample but from different vials/portions/analytes/aliquotes is averaged; data from different samples are combined into genomicMatrix; all data is then $\log_2(\text{fpkm}+1)$ transformed. The platform is Illumina. The type of data is gene expression RNAseq. The data link and descriptions are https://xenabrowser.net/datapages/?dataset=TCGA-BRCA.htseq_fpkm.tsv&host=https%3A%2F%2Fgdc.xenahubs.net&removeHub=https%3A%2F%2Fxcena.treehouse.gi.ucsc.edu%3A443.

Results and Interpretations

Using a probability higher than 50% as the threshold, we identify three critical DEGs: RBM22 (RNA binding motif protein 22), RNF213 (ring finger protein 213), and CACNG4 (Calcium Voltage-Gated Channel Auxiliary Subunit Gamma 4), which lead to 100% sensitivity and 100% specificity of classifying all 265 samples in their respective groups in the first TNBC dataset; four critical DEGs: MYCT1 (MYC Target 1), NUA2 (NUAK Family Kinase 2), NAT8L (N-Acetyltransferase 8 Like), and CACNG4, which lead to 100% sensitivity and 100% specificity of classifying all 66 samples in their respective groups in the second TNBC dataset; four critical DEGs: MYCT1, UNC5B (Unc-5 Netrin Receptor B), NUA2, and NAT8L, which also lead to 100% sensitivity and 100% specificity of classifying all 121 samples in their respective groups in the third breast cancer dataset; and the same four critical DEGs as in the third dataset, which leads to 100% sensitivity and 96.5% specificity of classifying all 1217 samples in their respective groups in the fourth breast cancer dataset.

Our final classifiers are combined classifiers of three competing factor ($CF_i, i=1,2,3$) classifiers expressed as:

For the first TNBC (North American cohort) dataset:

Data-1-CF1: 19.0107 +3.1105*RNF213 - 3.6692*CACNG4
Data-1-CF2: -0.4312 +8.0992*RNF213 - 9.5921*RBM22
Data-1-CFmax: max(Data-1-CF1, Data-1-CF2)

For the second TNBC (European cohort) dataset:

Data-2-CF1: 39.8651 -1.6945*NAT8L -3.5933*CACNG4
Data-2-CF2: 9.8676 -5.1333*MYCT1 +0.4595*NUAK2
Data-2-CFmax: max(Data-2-CF1, Data-2-CF2)

For the third (European cohort) dataset:

$$\text{Data-3-CF1: } 25.1089 - 10.1863 * \text{MYCT1} + 3.1654 * \text{NUAK2} - 2.0708 * \text{NAT8L}$$

$$\text{Data-3-CF2: } 2.4425 + 2.0119 * \text{UNC5B} - 4.1677 * \text{NUAK2} - 0.8255 * \text{NAT8L}$$

$$\text{Data-3-CFmax: } \max(\text{Data-3-CF1}, \text{Data-3-CF2})$$

For the fourth (Genomic Data Commons) dataset:

$$\text{Data-4-CF1: } 5.7644 - 2.5133 * \text{MYCT1} + 2.3383 * \text{NUAK2} - 1.2537 * \text{NAT8L}$$

$$\text{Data-4-CF2: } -9.5458 + 3.1219 * \text{UNC5B} + 0.7849 * \text{NUAK2}$$

$$\text{Data-4-CF3: } 7.0281 - 2.9389 * \text{MYCT1} + 4.3574 * \text{NUAK2} - 2.8591 * \text{NAT8L}$$

$$\text{Data-4-CFmax: } \max(\text{Data-4-CF1}, \text{Data-4-CF2}, \text{Data-4-CF3})$$

The risk probabilities are calculated using the logistic function of $\exp(\text{Data-}i\text{-CFmax}) / (1 + \exp(\text{Data-}i\text{-CFmax}))$ for the combined classifiers in each dataset, or $\exp(\text{Data-}i\text{-CF}j) / (1 + \exp(\text{Data-}i\text{-CF}j))$ for each individual classifier $i=1,2,3, 4, j=1,2,3$.

In the first three cohorts, multiple ID-ref subtype genes correspond to a gene symbol. The following ID-ref subtype genes are used in the classifiers: 236872_at (RBM22), 241480_at (RNF213), 62987_r_at (CACNG4), 220471_s_at (MYCT1), 220987_s_at (NUAK2), 228880_at (NAT8L), 226899_at (UNC5B).

Table 1. Gene information, expression values, competing factors, risk probabilities										
The first (TNBC) dataset										
ID	TNBC /No	236872_at	241480_at	62987_r_at		CF1	CF2		CFmax	Pmax
GSM1974566	1	6.09	7.42	10.55		3.37	1.19		3.37	0.97
GSM1974567	1	5.80	7.35	10.23		4.32	3.47		4.32	0.99
.....										
GSM1974763	1	5.82	7.25	10.20		4.15	2.52		4.15	0.98
GSM1978883	0	6.15	6.81	11.50		-2.01	-4.26		-2.01	0.12
.....										
GSM1978948	0	6.17	6.69	11.92		-3.93	-5.47		-3.93	0.02
GSM1978949	0	6.21	6.73	11.73		-3.11	-5.45		-3.11	0.04
The second (TNBC) dataset										
ID	TNBC /No	220471_s_at	220987_s_at	228880_at	62987_r_at	CF1	CF2		CFmax	Pmax
GSM1588970	1	2.37	5.26	5.47	8.08	1.58	0.13		1.58	0.83
GSM1588971	1	2.36	5.40	4.86	8.25	1.97	0.25		1.97	0.88
.....										
GSM1589150	0	2.42	4.98	8.83	8.64	-6.16	-0.26		-0.26	0.44
GSM1589151	0	2.46	5.02	8.95	8.57	-6.09	-0.47		-0.47	0.38
The third breast cancer dataset										

ID	BC /NoB C	220471_s_a t	226899_at	220987_s_at	228880_at	CF1	CF2		CFmax	Pmax
GSM1045191	0	3.21	6.95	2.94	5.65	-10.01	-0.49		-0.49	0.38
GSM1045192	0	2.90	7.88	3.57	4.27	-2.00	-0.11		-0.11	0.47
GSM1045193	0	3.16	6.57	3.19	4.67	-6.67	-1.49		-1.49	0.18
.....										
GSM1045207	0	2.31	6.68	3.41	7.61	-3.40	-4.63		-3.40	0.03
GSM1045208	1	2.31	8.56	4.02	4.43	5.08	-0.73		5.08	0.99
.....										
GSM1045310	1	2.31	7.84	5.54	4.67	9.40	-8.70		9.40	1.00
GSM1045311	1	2.31	9.91	4.09	4.85	4.43	1.35		4.43	0.99
The fourth breast cancer dataset										
ID	BC /NoB C	MYCT1	UNC5B	NUAK2	NAT8L	CF1	CF2	CF3	CFmax	Pmax
TCGA-E9-A1NI-01A	1	1.82	3.40	2.24	1.08	5.09	2.82	8.38	8.38	1.00
TCGA-A1-A0SP-01A	1	1.48	2.74	1.27	3.16	1.05	-0.01	-0.82	1.05	0.74
TCGA-BH-A1EU-11A	0	3.34	2.01	1.12	2.21	-2.79	-2.39	-4.23	-2.39	0.08
.....										
TCGA-BH-A0BW-11A	0	2.86	2.85	2.12	1.98	1.05	1.03	2.20	2.20	0.90
.....										
TCGA-BH-A0DK-11A	0	2.08	1.77	0.56	0.94	0.68	-3.57	0.69	0.69	0.67
.....										
TCGA-E2-A1LH-11A	0	2.27	1.92	1.46	2.16	0.76	-2.39	0.53	0.76	0.68
.....										
TCGA-AC-A2FF-11A	0	2.91	3.18	1.18	1.33	-0.46	1.31	-0.19	1.31	0.79
.....										
TCGA-A7-A5ZW-01A	1	2.42	3.96	1.52	0.42	2.69	4.01	5.31	5.31	1.00
TCGA-BH-A203-01A	1	1.74	2.87	1.44	0.79	3.77	0.54	5.94	5.94	1.00

Table 1 lists gene expression values, individual classifiers' computed values, the combined classifier's computed values, and the risk probabilities. Figure 1 plots all patients' risk probabilities with circles for breast cancer samples and asters for non-breast cancer samples. Figure 2 is a Venn diagram that plots individual classifiers' performance. This study is the first time TNBC and other breast cancer types can be further classified into subtypes based on critical genes' functions. This

new classification opens a new research direction, new drug developments, and new refined personalized therapies.

For the first TNBC (North American cohort) dataset, three genes (RNF213, RBM22, CACNG4) completely classify all 198 TNBC tumor samples into three subtypes (Figure 2) with the sensitivity of 100% and the specificity of 100%. From the individual classifiers, we can see that a decrease of RNF213 level will reduce the risk of developing TNBC, while increases in the expression levels of RBM22 and CACNG4 will reduce the risk of developing TNBC.

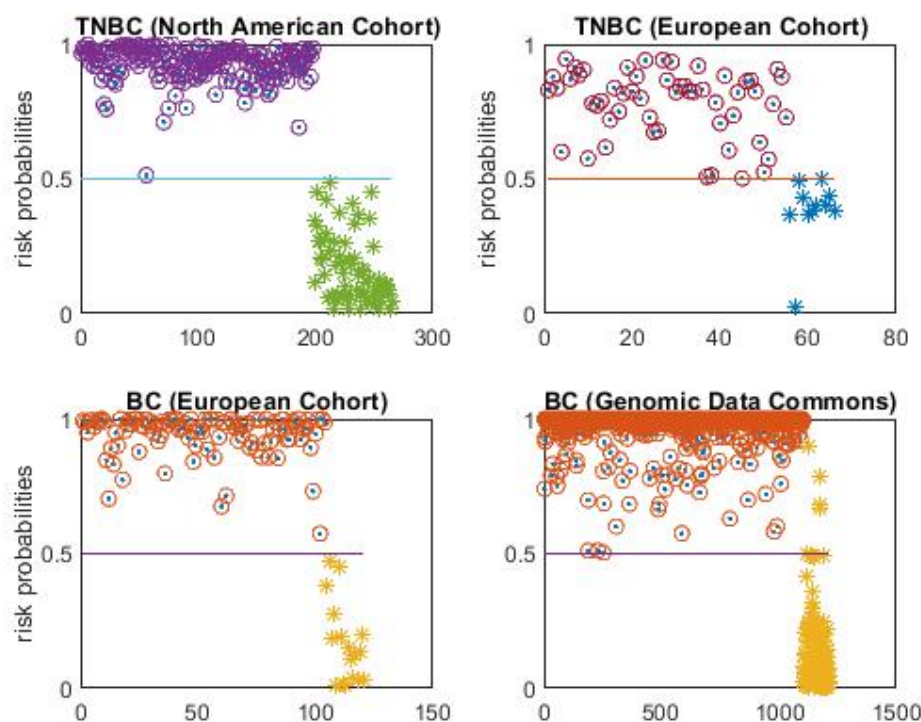


Figure 1. Risk probabilities of four cohorts. The circles are for patients with breast cancers. The asters are for tissues without breast cancers.

For the second TNBC (European cohort) dataset, four genes (MYCT1, NAT8L, NUAK2, CACNG4) completely classify all 66 TNBC tumor samples into three subtypes (Figure 2) with the sensitivity of 100% and the specificity of 100%. From the individual classifiers, we can see that a decrease in NUAK2 level will benefit the patients, while increases in the expression levels of MYCT1, NAT8L, and CACNG4 will benefit the patients. We note that there are also Her2, Luminal A and Luminal B samples in this second dataset. After adding classifier CF3: $21.8170 - 8.8170 \cdot \text{RBM22} - 0.3047 \cdot \text{NAT8L}$, all breast cancer (TNBC, Her2, Luminal A, Luminal B) patients will again be 100% accurately classified into their respective groups.

Comparing the first and second TNBC cohorts, we see that the TNBC patients from North American and the TNBC patients from European cohorts share a common gene CACNG4 and similar coefficients (-3.6692 vs. -3.5933). Otherwise, other critical genes from these two cohorts are different. This observation tells that the causes, the formations, and the therapies of TNBC can be

different from region to region and race to race. We want to note that based on our knowledge in the field, there does not exist any other method that can 100% accurately classify breast cancer patients and cancer-free patients into their respective groups. With 100% accuracy, regardless of how big and how small the sample is, these genes should contain basic cancer information of TNBC disease, they should be thoroughly analyzed and explored.

On the other hand, cautions should be called with any other classifiers with lower accuracy. Using genes derived/obtained from low accuracy classifiers may lead to suboptimal results and even wrong conclusions. The formulas of these two cohorts disclose the puzzle of TNBC as gene-gene interactions and functional effects are different. Such differences can be the most important part of studying TNBC and point out new research directions for better understanding TNBC and designing better treatments.

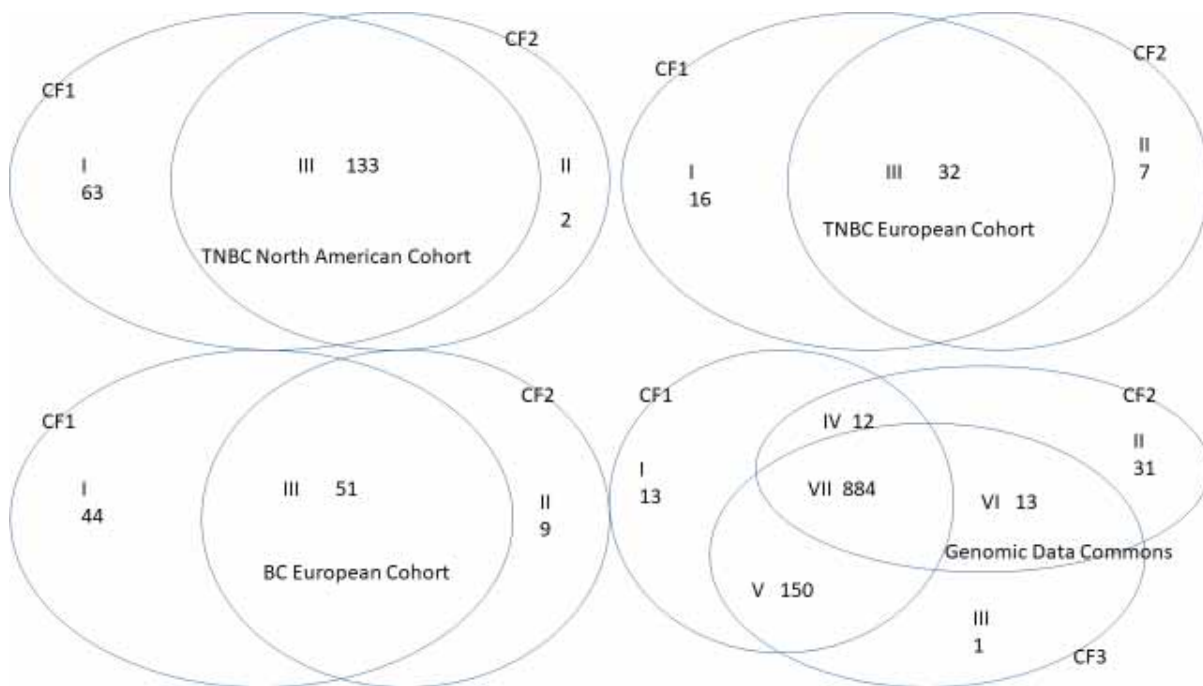


Figure 2. Venn diagrams of breast cancer subtypes. The first three cohorts have more than three subtypes. The fourth cohort has more than seven subtypes.

For the third (European cohort) dataset, four genes (MYCT1, NAT8L, NUA2, UNC5B) completely classify all 104 tumor samples into three subtypes (Figure 2) with a sensitivity of 100% and a specificity of 100%. A decrease of UNC5B level will benefit the patients in this cohort, while increases of expression levels of MYCT1 and NAT8L will benefit the patients. In addition, it can be seen that NUA2 can benefit the patients and can also harm the patients depending on the patients' breast cancer subtypes in Figure 2. These gene-gene relationships and genes-subtypes relationships tell efficient therapies to breast cancer patients depending on their subtypes' determinations.

For the fourth (Genomic Data Commons) dataset, the same four genes (MYCT1, NAT8L, NUA2, UNC5B) as for the third (European cohort) dataset completely classify 1104 tumor samples into

seven subtypes (Figure 2) with the sensitivity of 100% and the specificity of 96.5%. There are 4 samples among 103 normal samples being classified as tumor samples. Note that this dataset does not offer multiple ID-ref subtypes. If genes' expression values are taken the same as those ID-ref subtypes, the specificity may be improved to 100%. In this cohort, increases of MYCT1 and NAT8L levels can benefit the patients, while decreases of UNC5B and NUAK2 levels will benefit the patients.

Comparing the third and fourth breast cancer cohorts, the individual classifiers Data-3-CF1, Data-4-CF1, and Data-4-CF3 have the same component genes and coefficient signs. Data-3-CF2 has one more gene, NAT8L, than Data-4-CF2. However, the signs of NUAK2 coefficients in these two individual classifiers are different. We further note that to have two similar individual classifiers Data-4-CF1 and Data-4-CF3 in the final classifier is completely new in machine learning literature. These observations further reveal that breast cancer formations are more complicated than simply looking at some high/low expression values of individual genes as in the literature. The most important relations in finding critical genes linked to breast cancers are gene-gene interactions, genes-individual classifiers interactions, and their functional effects.

Comparing the second TNBC cohort, the third breast cancer cohort, and the fourth cohort, we see that increasing the levels of MYCT1 and NAT8L can benefit all patients.

In Figure 2, Venn diagrams for four different cohorts are different, with two TNBC cohorts having similar patterns, while BC European cohort and Genomic Data Commons are different. It is because the numbers of component classifiers in the final classifiers for different cohorts are different. Such phenomena tell that there are commonalities among breast cancer patients and specificities from patient to patient, i.e., the critical cancer informatics are expressed. Note that in Venn diagrams, the more intersections the groups, the more complex the disease, and the more difficult the treatment. Taking Genomic Data Commons as an example, patients in Group VII will be the most difficult to treat.

Figure 3 presents the gene-gene interactions, gene-subtype interactions, and functional effects of our identified competing classifiers. We can see clear signature patterns in each plot. This visualization tool provides a new way for breast cancer diagnosis.

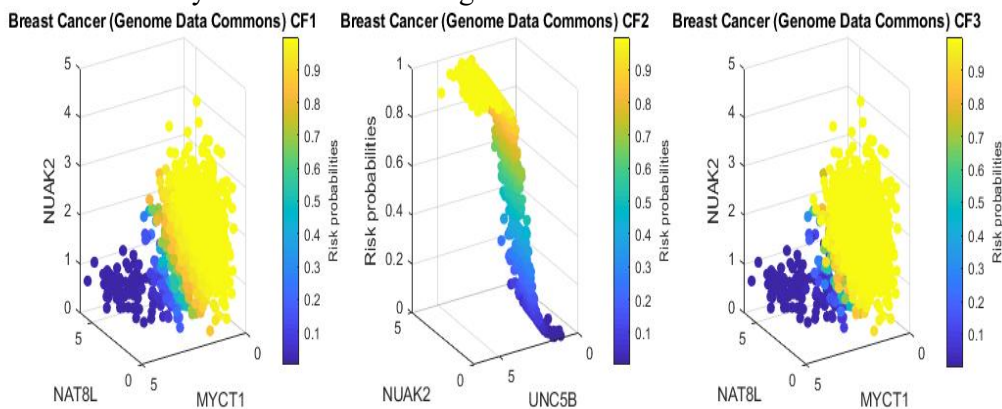


Figure 3. Four-dimensional plots for visualizing risk signature patterns from three competing component classifiers and the combined functional effects of gene-gene interactions and gene-subtype interactions of four genes.

Characteristics of studying samples

All four datasets are accompanied by some characteristics of patients. Here we report their inter-relationship with the competing classifiers. Table 2 displays Sex, Age, BMI, and Grade from the first dataset (TNBC, a North American cohort). Table 3 displays Age and BMI from the second dataset (TNBC samples only, A European cohort). Table 4 is for the third dataset (breast cancer samples, A European cohort). Finally, Table 5 includes disease Stage besides Age and Sex for the fourth breast sample data set (A Genomic Data Commons – TCGA).

	Sex		Age				BMI			Grade		
	Male	Female	<=50	(50,60]	(60,70]	>70	Normal	Overweight	Obese	Poorly	Moderately	Well
CF-1	0	63	34	12	10	5	13	19	61	33	25	1
CF-2	0	2	1	0	1	0	0	1	2	1	0	0
CF-(1,2)	0	133	51	33	28	18	36	41	129	74	28	3

	Age				BMI		
	<=50	(50,60]	(60,70]	>70	Normal	Overweight	Obese
CF-1	3	4	2	2	5	3	3
CF-2	2	2	0	1	2	2	1
CF-(1,2)	9	10	5	1	12	6	6

	Age				Grade		
	<=50	(50,60]	(60,70]	>70	Poorly	Moderately	Well
CF-1	9	15	11	9	25	14	5
CF-2	3	4	2	0	4	5	0
CF-(1,2)	15	13	10	13	24	21	6

	Age					Sex		Stage				
	<=50	(50,60]	(60,70]	(70,80]	>80	Male	Female	I	II	III	IV	X
CF-1	8	1	3	0	1	0	13	4	6	3	0	0
CF-2	10	9	7	2	1	1	30	6	16	9	0	0
CF-3	1	0	0	0	0	0	1	0	0	1	0	0
CF-(1,2)	5	1	4	1	1	0	12	2	7	2	0	1
CF-(1,3)	44	36	35	27	7	2	148	25	97	24	2	0
CF-(2,3)	2	5	5	1	0	0	13	3	3	5	1	1
CF-(1,2,3)	257	216	225	119	43	9	862	143	490	202	17	10

Overall, these four tables show that more patients fall in groups related to more than one competing classifier. Obese patients can make TNBC more complex. The more the competing classifiers, the worse the grade. In Table 5, Stages (IV, X) are mainly related to CF-(1,2,3), which shows the classifiers are positively correlated.

Discussions

This study is the first time in the medical literature that breast cancer diseases can be classified almost 100% correctly using only a few (three or four) genes. The results clearly disclose the puzzle of breast cancers, including TNBC, due to the selected genes and their predicting powers through gene-gene interaction, gene-subtype interaction, and functional effects. The results also point to new treatment directions.

We note that this study does not use the primary endpoint information. It is a pure classification study. The main purpose is to identify the essential breast cancer informatics. The study has achieved 100% accuracy, which is the first in the literature. Given patients have different endpoints, the new classifier still reaches 100% accuracy, which means the classifier is robust to patients' disease states, i.e., we can conclude that the classifier is robust regardless of primary endpoints and other individual attributes.

The discovery of critical genes can motivate many new research directions and laboratory experiments. These critical genes and their derived signature patterns (individual classifiers) can be a starting point as new biomarkers for conducting gene network analysis, testing other reported genes, and finding the causal directions of gene expression in various projects. As a result, many other existing pieces of research can be enriched. It can also be hoped that new types of diseases can be discovered. Eventually, new testing procedures and therapies for breast cancer can be designed.

These critical genes enrich the biological literature of their new functions related to breast cancer from indirect relationship to direct relationship. In many scenarios, indirect effects are more significant than direct effects as direct effects can be seen and controlled, while indirect effects are hard to see and even not to say how to control.

The risk probability of a patient developing a specific type of breast cancer in her/his life is low. Among all discovered breast cancer types, growing more than one type of breast cancer is rare. These breast cancer types compete, and one type will first be diagnosed. As a result, the competing risk factor models can efficiently model multiple breast cancer types.

This study's inference/analysis approach can shed new light on all gene-related research, i.e., not just the breast cancer study. Researchers can apply max-linear type models in their studies. Ultimately, our new findings may make researchers' cancer research efforts more effective and meaningful, reduce substantial research costs, and save lives and protect people.

Finally, we address an important medical practice issue. In this paper, all classifier formulas are explicitly expressed. Thus, the results in Table 1 are reproducible. Furthermore, Figures 1 and 3 show the risks of all patients. Using this paper's results, medical doctors have a powerful tool (testing kit) in their daily work, i.e., in the diagnostic stage, diagnosing and analyzing patients' breast cancer risks based on the four or fewer critical genes' expression values and the computed risks; in the treatment stage, those signature patterns can be used to study the effectiveness of drugs and treatments, i.e., conduct clinical trials, e.g., survival analysis, based on classified groups; in the drug development stage, pharmaceutical companies can use the findings of critical genes to study new drugs; finally, it can be hoped that mRNA-based therapies can be introduced using the critical genes' information in the therapy stage.

References

- Amjad, E., Asnaashari, S., Sokouti, B. et al. Systems biology comprehensive analysis on breast cancer for identification of key gene modules and genes associated with TNM-based clinical stages. *Sci Rep* 10, 10816 (2020). <https://doi.org/10.1038/s41598-020-67643-w>
- Breast Cancer Association Consortium. Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. *N Engl J Med* 2021; 384:428-439 DOI: 10.1056/NEJMoa1913948
- Burstein MD, Tsimelzon A, Poage GM, Covington KR et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res* 2015 Apr 1;21(7):1688-98. PMID: 25208879
- Chen H, Wu J, Zhang Z, et al. Association Between *BRCA* Status and Triple-Negative Breast Cancer: A Meta-Analysis. *Front Pharmacol.* 2018;9:909. Published 2018 Aug 21. doi:10.3389/fphar.2018.00909
- Clarke C, Madden SF, Doolan P, Aherne ST et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* 2013 Oct;34(10):2300-8. PMID: 23740839
- Cui, Q., Xu, Y., Zhang, Z., et al., Max-linear regression models with regularization. *Journal of Econometrics* (2020). Available online 1 August. <https://doi.org/10.1016/j.jeconom.2020.07.017>
- Cui, Q., Zhang, Z., Max-linear competing factor models. *Journal of Business and Economic Statistics* (2018), **36**, 62-74. <http://amstat.tandfonline.com/doi/full/10.1080/07350015.2015.1137761>
- Deng J.-L., Xu, Y.-H., Wang, G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. *Frontiers in Genetics.* (2019), 695, <https://www.frontiersin.org/article/10.3389/fgene.2019.00695>
- Dong P, Yu B, Pan L, Tian X, Liu F. Identification of Key Genes and Pathways in Triple-Negative Breast Cancer by Integrated Bioinformatics Analysis. *Biomed Res Int.* 2018 Aug 2;2018:2760918. doi: 10.1155/2018/2760918. PMID: 30175120; PMCID: PMC6098886.
- den Hollander P, Rawls K, Tsimelzon A, Shepherd J et al. Phosphatase PTP4A3 Promotes Triple-Negative Breast Cancer Growth and Predicts Poor Patient Survival. *Cancer Res* 2016 Apr 1;76(7):1942-53. PMID: 26921331
- Hu, C., Hart, SN, Gnanaolivu, R. et al. A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med* 2021; 384:440-451 DOI: 10.1056/NEJMoa2005936
- Lin Y, Fu F, Lv J, et al. Identification of potential key genes for HER-2 positive breast cancer based on bioinformatics analysis. *Medicine (Baltimore).* 2020;99(1):e18445. doi:10.1097/MD.00000000000018445
- Lu X, Gao C, Liu C, Zhuang J, Su P, Li H, Wang X, Sun C. Identification of the key pathways and genes involved in HER2-positive breast cancer with brain metastasis. *Pathol Res Pract.* 2019 Aug;215(8):152475. doi: 10.1016/j.prp.2019.152475. Epub 2019 Jun 1. PMID: 31178227.
- Lv, X., He, M., Zhao, Y. et al. Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer. *Cancer Cell Int* 19, 172 (2019). <https://doi.org/10.1186/s12935-019-0884-0>
- Malinowski, A., Schlather, M., and Zhang, Z., Intrinsically weighted means and non-ergodic marked point processes. *Annals of the Institute of Statistical Mathematics* (2016), **68**, 1-24.
- Malvia, S., Bagadi, S.A.R., Pradhan, D. et al. Study of Gene Expression Profiles of Breast Cancers in Indian Women. *Sci Rep* 9, 10018 (2019). <https://doi.org/10.1038/s41598-019-46261-1>
- Maubant S, Tesson B, Maire V, Ye M et al. Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. *PLoS One* 2015;10(4):e0122333. PMID: 25848952
- Maire V, Baldeyron C, Richardson M, Tesson B et al. TTK/hMPS1 is an attractive therapeutic target for triple-negative breast cancer. *PLoS One* 2013;8(5):e63712. PMID: 23700430

- Maire V, Némati F, Richardson M, Vincent-Salomon A et al. Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. *Cancer Res* 2013 Jan 15;73(2):813-23. PMID: 23144294
- Narod , SA, Which Genes for Hereditary Breast Cancer? *N Engl J Med* 2021; 384:471-473 DOI: 10.1056/NEJMe2035083
- Xu, Y., Regression models with max-linear structure. *Ph.D. Dissertation* (2019). University of Wisconsin at Madison.
- Zhang, Z., Dimension reduction from ultra-high to ultra-low: A three-variable selector - The crazy, the good, and the ugly. *International Conference on High Dimensional Data Analysis, Kunming, China* (2007); Invited talk.
- Zhang, Z., Quotient correlation: a sample-based alternative to Pearson's correlation, *Annals of Statistics* (2008), **36**, 1007-1030.
- Zhang, Z., On studying extreme values and systematic risks with nonlinear time series models and tail dependence measures (with discussion). *Statistical Theory and Related Fields* (2020). In print 10.1080/24754269.2020.1856590.
- Zhang, Z., Rejoinder of "On studying extreme values and systematic risks with nonlinear time series models and tail dependence measures" *Statistical Theory and Related Fields* (2021). <https://doi.org/10.1080/24754269.2021.1871710>
- Zhang, Z., Five critical genes related to seven Covid-19 subtypes. (2021). *Journal of Data Science*, Volume 19, Issue 1, pp. 142–150. <https://doi.org/10.6339/21-JDS1005>
- Zhang, Z., Zhang, C., and Cui, Q., Random threshold driven tail dependence measures with application to precipitation analysis. *Statistica Sinica* (2017), **27**, 685-709.
- Zhong G, Lou W, Shen Q, Yu K, Zheng Y. Identification of key genes as potential biomarkers for triple-negative breast cancer using integrating genomics analysis. *Mol Med Rep.* 2020;21(2):557-566. doi:10.3892/mmr.2019.10867
- Zhu C, Ge C, He J, Zhang X, Feng G, Fan S. Identification of Key Genes and Pathways Associated With Irradiation in Breast Cancer Tissue and Breast Cancer Cell Lines. *Dose Response.* 2020 Jun 18;18(2):1559325820931252. doi: 10.1177/1559325820931252. PMID: 32684870; PMCID: PMC7346585.