Journal of King Saud University – Science 35 (2023) 102447

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: www.sciencedirect.com

Original article

Gene expression study of breast cancer using Welch Satterthwaite *t*-test, Kaplan-Meier estimator plot and Huber loss robust regression model

Sajjad Karim [a,b,1,*], Md Shahid Iqbal [c,1], Nesar Ahmad [c], Md Shahid Ansari [d], Zeenat Mirza [b,e], Adnan Merdad [f], Saddig D. Jastaniah [g], Sudhir Kumar [h]

[a] Center of Excellence in Genomic Medicine Research, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
[c] University Department of Statistics and Computer Applications, T. M. Bhagalpur University, Bhagalpur, India
[d] Department of Clinical Data Analytics, Max Super Speciality Hospital, New Delhi, India
[e] King Fahd Medical Research Center, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
[f] Surgery Department, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia
[g] Department of Diagnostic Radiology, Faculty of Applied Medical Science, King Abdulaziz University, Jeddah, Saudi Arabia
[h] Institute for Genomics and Evolutionary Medicine, and Department of Biology, Temple University, Philadelphia, PA 19122, USA

ARTICLE INFO

Article history:
Received 21 May 2022
Revised 21 October 2022
Accepted 12 November 2022
Available online 17 November 2022

Keywords:
Breast cancer
Gene expression
Microarray
Welch Satterthwaite *t*-test
Kaplan-Meier plot
Huber loss robust regression

ABSTRACT

Objective: Breast Cancer (BC) is one of the deadliest diseases in women, causing thousands of deaths annually despite the advent of high-throughput genomic platforms in the recent past. Microarray-based gene expression profiling with different statistical methods have been extensively used to understand the disease at the molecular level. We plan to apply Welch Satterthwaite *t*-test, Kaplan-Meier estimator plot and Huber Loss robust regression model on microarray data to improve the analysis and find biomarkers for future diagnosis, prognosis, and treatment.

Methods: We retrieved microarray data (GSE10810 dataset) of 31 breast tumor samples and 27 normal breast samples from Gene Expression Omnibus (GEO, NCBI). Welch Satterthwaite *t*-test was applied to identify the most statistically significant genes, Huber loss robust regression model was applied to investigate the existing mathematical relations between tumor and control variables, and Kaplan-Meier Plotter was used to confirm their association with overall metastatic relapse-free survival of BC patients.

Results: We identified 1837 differentially expressed genes, including 638 overexpressed (*COL11A1, KIAA0101, S100P, GJB2, TOP2A, LINC01614, RRM2, INHBA, C15orf48* and *CKS2*) and 1199 under expressed (*LEP, ADIPOQ, PLIN1, PCK1, PCOLCE2, ADH1B, LYVE1, FABP4, ABCA8,* and *CHRDL1*) genes passing the threshold (fold change ± 2 and p value < 0.001). KM analysis revealed 12 out of 20 DEGs (log rank p value < 0.05) as potential prognostic and therapeutic biomarkers.

Conclusion: Huber loss robust regression model was found to be one of the best performing algorithms for the mathematical relationship between the control and breast tumor samples with co-relation coefficient of 0.4398 and mean absolute error of 1.069 ± 0.020. In conclusion, with high mathematical confidence, we detected DEGs have high potential to be BC biomarkers using Welch *t*-test and Kaplan-Meier plot having minimum underlying assumptions.

* Corresponding author at: Center of Excellence in Genomic Medicine Research, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.
E-mail addresses: skarim1@kau.edu.sa (S. Karim), shahid273@gmail.com (M.S. Iqbal), ahmad_n@tmbuniv.ac.in (N. Ahmad).
[1] Co-first authors contributed equally.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksus.2022.102447
1018-3647/

1. Introduction

Cancer is a complex disease where irregular cell differentiation and proliferation converts normal cells into tumors. Individual's genetic factors stimulated by carcinogenic factors cause cancer (Parkin, 2006; Plummer et al., 2016; Zapatka et al., 2020). In 2020, the World Health Organization reported 1.8 million deaths from lung cancer, 935,000 deaths from colorectal cancer, 830,000 deaths from liver cancer, 769,000 deaths from stomach cancer, and 685,000 deaths from breast cancer (BC). BC usually affect the epithelium of the ducts (85 %) or lobules (15 %) in the glandular tissue of the breast ("Breast cancer", Who.int, 2021). *BRCA1* and *BRCA2* genes are frequently used as an inherited diagnostic marker. However, hundreds of genes and pathways have been found to be associated with BC. Therefore, a detailed functional study is needed to understand the complexity and polymorphisms of cancer at the genetic level.

Recent advent of genomic and trancriptomic technologies have helped researchers to find variation at the nucleotide level and determine the simultaneous expression of thousands of genes at any specific stage of BC (Russo et al., 2003). The selection of the most appropriate statistical methods/models is a key step in microarray data analysis to identify the significantly associated up-and down-regulated genes with a higher level of confidence. Mathematical model like Pearson's correlation is used to measure the relation between gene expression values for linearly associated data, whereas rank correlation is preferred for nonlinear data (de Siqueira Santos et al., 2013). Student *t*-test is a commonly used statistical method for comparing two independent groups in clinical data that might give biased results because of the underlying assumption of normality and homoscedasticity (homogeneity of variance), and lead to unsound and unreliable mathematical inferences (Erceg-Hurn and Mirosevich, 2008). Welch Satterwaite's *t*-test, Yuen's *t*-test, and a bootstrapped *t*-test are other popular *t*-tests based on the underlying assumption and used for analysis (Rasch et al., 2009; Delacre et al., 2017).

We aim to check the mathematical relation between tumor and control. Outliers are the troublemaker while applying any statistical model to determine the mathematical relation. We compared the efficiency results of linear, Huber, RANSAC, and Theil-Sen robust regression models and used Huber loss robust regression model to investigate the mathematical correlation between tumor and control samples.

Cross-validation of DEGs using qPCR brings confidence in high-throughput result. Prognostic values of genes could be determined by survival probability using Kaplan Meier (KM) survival estimator, Nelson–Aalen estimator, Cox Proportional Hazard Model based on regression (Kaplan and Meier, 1958; Cox, 1972).

In the present study, we used a microarray dataset for (i) re-analyzing the experiment to identify key differentially expressed genes, (ii) validating survival associated with most altered genes using web-based Kaplan Meier Plotter tool, and (iii) investigating the existence of potential mathematical relationships between tumor and control variables.

2. Materials and methods

2.1. Data collection:

We obtained gene expression microarray raw data as.CEL files of "GSE10810" dataset from Affymetrix Human Genome U133 Plus 2.0 Array [GPL570] with 54,675 probes (Pedraza et al., 2010). The cohort contains 58 samples including 31 BC and 27 control.

2.2. Welch Satterthwaite t-test for identification of differentially expressed genes

The Welch Satterthwaite *t*-test was applied to compare the mean of control and BC samples and to detect the significant difference between control and tumor groups using the following formula:

$$\omega(t) = \frac{\Delta\bar{X}}{s_{\Delta\bar{X}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}} \tag{1}$$

$$s_{\bar{X}^-{}_i} = \frac{s_i}{\sqrt{Ni}} \tag{2}$$

Here, $\bar{X}_i$ = i^{th} sample mean.

$s_{\bar{X}_i}$ = i^{th} standard error, for a given sample standard deviation and sample size, the denominator is not primarily linked with pooled variance estimate.

The degrees of freedom: Welch degree of freedom = $\omega(v)$ combined with this variance estimate, is approximated using the Welch–Satterthwaite equation

$$\omega(v) \approx \frac{\left(\frac{s^2_1}{N_1} + \frac{s^2_2}{N_2}\right)^2}{\frac{s^4_1}{N^2_1\omega(v_1)} + \frac{s^4_2}{N^2_2\omega(v_2)}} \tag{3}$$

In case of $N_1 = N_2$

$$\omega(v) \approx \frac{s^4{}_{\Delta\bar{X}}}{\omega(v_1)^{-1}s^4{}_{\bar{X}_1} + \omega(v_2)^{-1}s^4{}_{\bar{X}_2}} \tag{4}$$

$\omega(v_i) = N_i - 1$ is the degree of freedom.

If the sample size and variance are equal then both Student *t*-test and Welch *t*-test behave same, however, changes with variance and sample size (Baguley, 2012). Based on cut-off *p*-values $\leq$0.05 and fold change ± 1.5, the model could be several stringent, moderate, and liberal that can give different results. Welch's *t*-test was applied on each row of 3126 probes for filtration and identification of significant differentially expressed genes.

2.3. Kaplan–Meier estimator for survival analysis:

The Kaplan-Meier estimator is a non-parametric model used for survival probability function with minimal assumptions. We assume the event takes place at a specific time, all the data points and censored observations have the same chance of surviving.

The Kaplan–Meier (KM) estimator (Kaplan and Meier, 1958) is mathematically expressed as:

$$SFatt = \prod_{i=t_i<t} \frac{n_i - d_i}{n_i} = \prod_{i=t_i<t} (1 - d_i/n_i) \tag{5}$$

SF = Survival Function.

n_i = number of people at risk at any given time t_i and d_i = the number of events occurring at any given time t_i

The survival curve remains constant between two occurrences, such as t_i andt_i + 1. Equation (5) can be rewritten using a recursive formula.

$$SFatt_j = \left[\frac{n_{j-1} - d_{j-1}}{n_{j-1}}\right] \text{ multiply by (SF) at } t_{j-2} \tag{6}$$

We used "Kaplan-Meier Plotter" to see if the expression levels of the selected up and down regulated genes were correlated to BC patient's, overall metastatic relapse-free survival with 95 % con-

fidence interval, calculated hazard ratio (HR), statistical significance log rank p value was ≤ 0.05 (https://kmplot.com/analysis/) (Emmert-Streib and Dehmer, 2019; Lánczky and Győrffy, 2021).

2.4. Huber loss robust regression model for mathematical correlation:

We applied the Huber loss robust regression model to investigate the mathematical correlation between BC and control samples. This model intends to minimize residuals and utilize the concept of loss function to precisely determine the expected outcome. Thus, it is critical to pick the best-fitting loss function [mean square error (MSE) and mean absolute error (MAE)] with certain weight to outliers (Gupta et al., 2019). MSE is the sum of the squared distances between the target variable and predicted values, and MAE is the sum of the absolute differences between our target and predicted variables:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - y_i^p)^2}{n} \tag{7}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i{}^p|}{n} \tag{8}$$

We used Huber loss/smooth mean absolute error, a mixture of both MSE and MAE. Huber loss is sensitive to outliers, differentiable at zero, the error becomes quadratic for small errors. Quadratic values depend upon the hyperparameter (δ, delta).

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}\left(y - f(x)\right)^2 for |y - f(x)| \leq \delta, \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 otherwise \end{cases} \tag{9}$$

We also compared the mean performance of each method, Linear Regression, Huber Regression, RANSAC Regression, and Theil-Sen Regression and used a box and whisker plot to compare the distribution of scores across the cross-validation folds.

2.5. Validation of microarray results by quantitative PCR:

We validated the expression of over-expressed (*KIAA0101, S100P, TO2A, RRM2, INHBA*) and under-expressed (*ADIPOQ, PLIN1, ADH1B, ABCA8, CHRDL1*) genes by qPCR assay using Applied Biosystems StepOnePlus Real-Time PCR instrument (ThermoFisher Scientific, USA). Quantification was performed using PowerUp™ SYBR™ Green Master Mix using *GAPDH1* as reference. DataAssist™ Software were used for initial Ct values calculation and comparative Ct (ΔΔCt) method was used for quantitative gene expression.

3. Results

A total of 54,675 probes mean were used to compare the expression values of tumors and controls with descriptive statistical parameters including mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, and range (Table 1). Initial analysis revealed 3126 probes passing the threshold values of fold change (±2) and p value (<0.05). We finally identified 1837 differentially expressed (up-and down-regulated) genes by applying Welch t-test at $p < 0.001$ and cross-validated discovered top 10 up and down expressed genes through KM survival analysis (**Supplementary Table 1**). A significant difference between tumor and control samples were established with the following values:

The mean value of tumor (6.824 ± 1.649) and control (7.41 4 ± 2.007), Welch's t value (-12.70), Welch-Satterthwaite degree of freedom (6022.8) and p value < 0.0001. For the normality assumption, control samples were considered as an independent variable, whereas tumor as a dependent variable and 3126 probes mean gene expression values were found not to be fit within the normality assumption as tested by "Shapiro-Wilk test" and "D'Agostino's K- squared" test. However, further investigation of the complete data set of 54,675 probes revealed a close to normal distribution of data and represented as histogram, probability- probability (PP), and quantile–quantile (QQ) plots for tumor and control samples (Fig. 1). Majority of the data set fits in the normal distribution while stragglers and curvature at either end of the normal probability indicated the lack of symmetry or presence of outliers in the dataset. We found skewness and kurtosis values between − 1 and + 1 indicating close to normal distribution.

Table 1
Comparison of expression values of tumor and control using descriptive statistics of 54,675 probe mean.

Descriptive Statistics	Mean Tumor	Mean Control
Mean	5.5674	5.5715
Standard Error	0.0085	0.0087
Median	5.1751	5.1403
Mode	5.205	5.4836
Standard Deviation	1.9788	2.0278
Sample Variance	3.9158	4.1121
Kurtosis	0.2685	0.274
Skewness	0.8154	0.8411
Range	11.3393	11.4354
Minimum	2.639	2.6563
Maximum	13.9782	14.0917

Furthermore, the Kaplan–Meier plot refined the experiment and analysis for the top 10 up- and-down-expressed genes as a drill-down approach and down streaming for survival analysis. Finally, we have 7 out of 10, *KIAA0101, S100P, TOP2A, RRM2, INHBA, C15orf48, and CKS2* as important up-expressed genes, while 5 out of 10, *ADIPOQ, PLIN1, ADH1B, ABCA8, and CHRDL1* are important down-expressed genes that can be considered as diagnostic, prognostic, and therapeutic biomarkers. Thus, in the present study, we selected the top 10 up and down regulated DEGs for discussion (Fig. 2).

Validation of differentially expressed genes were performed by real time PCR (qPCR) by calculating mean Rq, fold change and p-values. The qPCR confirmed the overexpression of *KIAA0101, S100P, TO2A, RRM2, INHBA* and under-expression of *ADIPOQ, PLIN1, ADH1B, ABCA8, and CHRDL1* in the BC tissues (Fig. 3).

Kaplan-Meier Plotter was used to confirm survival in a larger dataset and its association with the identified genes. KM Plot (Figs. 4 and 5) displays the top 9 differentially up and down-expressed genes, their Hazard ratio with 95 % confidence interval, log rank p values (Table 2**)**. We consider the gene a significant biomarker for prognostic and therapeutic importance if the log rank p value < 0.05.

Based on the Huber loss robust regression model, a weak correlation (0.439824) between the control and BC samples was found, representing significant difference in the two groups (cancer and control) as expected. Mean absolute error for Linear, Huber, RANSAC and Theil Sen Regression were 1.075 ± 0.020, 1. 069 ± 0.020, 1.245 ± 0.105 and 1.093 ± 0.018 respectively. Comparative analysis results revealed Huber as the best performing regression model with MEA with standard deviation = 1.069 ± 0. 020. A box and whisker plot revealed the distribution of results for each evaluated algorithm and lower distributions for the Huber robust regression algorithm found in compared to other linear regression algorithms. We have also shown best-fit line equations through linear Huber loss robust regression model, RANSAC Regression, and Theil Sen regression model (Table 3**,** Fig. 6**)**.

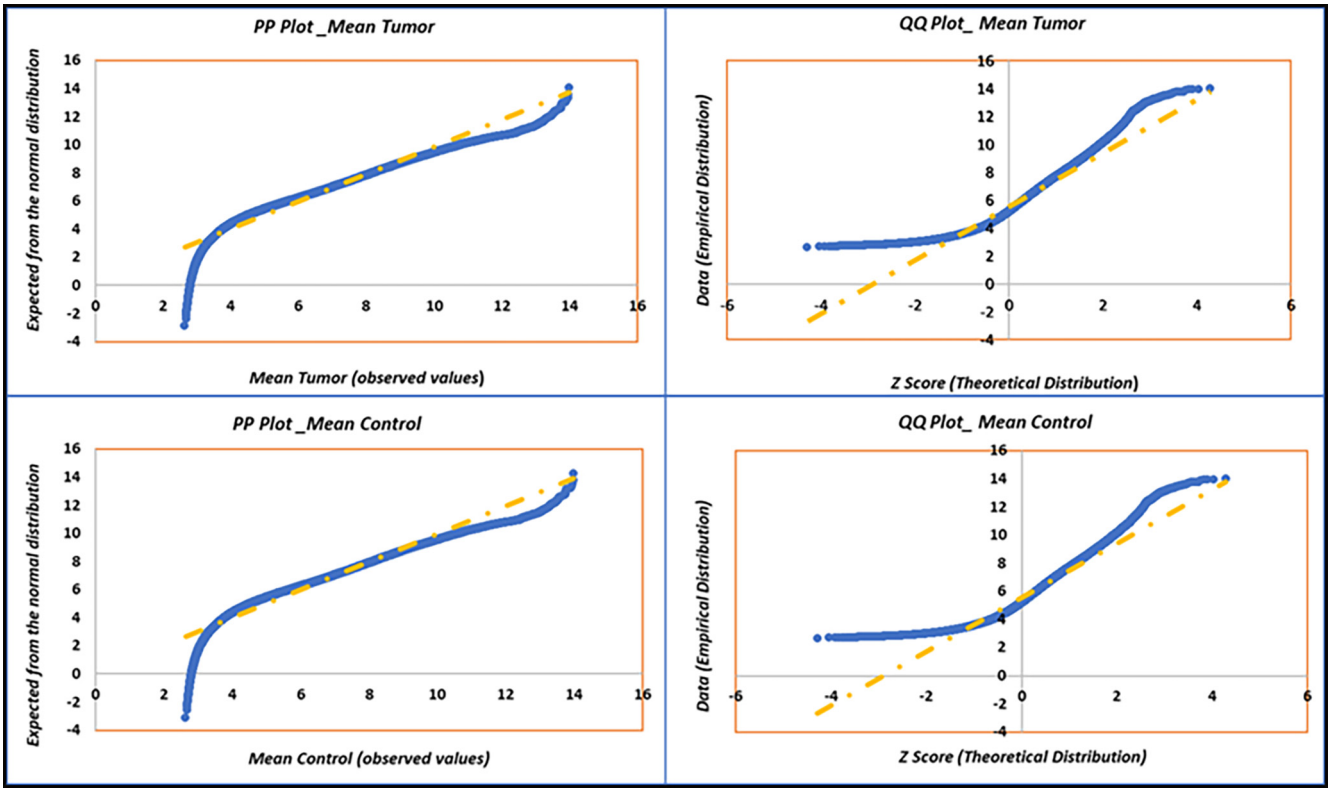

Fig. 1. PP (Probability-Probability) and QQ (Quantile-Quantile) plots for the tumor and control variables with 54,675 probes mean gene expression values for normality check as the underlying assumption of Welch *t*-test: **(A(top left))** PP plot of tumor, **(B(bottom left))** PP plot of control, **(C(top right))** QQ plot of tumor, and **(D(bottom right))** QQ plot of control.

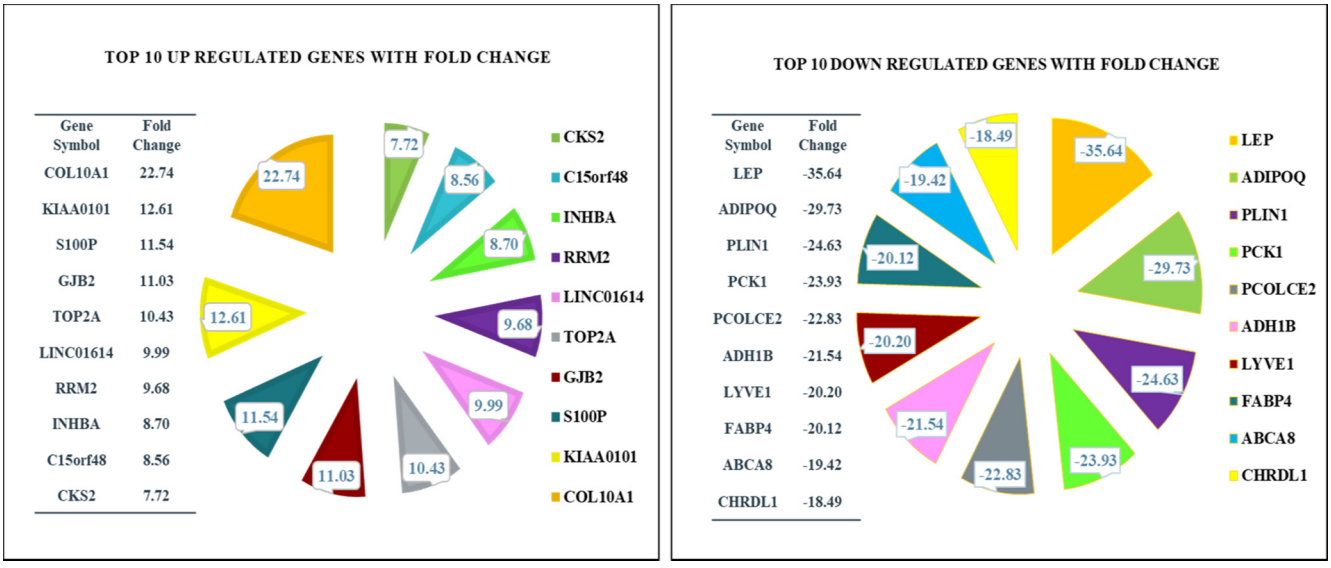

Fig. 2. Graph represents the fold change of the top 10 up- and down-regulated differentially expressed genes after passing the filtration criteria.

4. Discussion

The aim of the present study was to search for precise and robust statistical methods to identify the differentially expressed genes with a higher degree of mathematical confidence. Student's *t*-test, Welch's *t*-test, Trimmed Means *t*-Test, Yuen-Welch's *t*-Test, and bootstrapped *t*-test are commonly used to compare two independent groups. Student's *t*-test, frequently used for clinical datasets, must fulfill the underlying assumption of normality and homoscedasticity (homogeneity of variance) as prerequisites. Violation of assumptions may lead to biased, unsound, and unreliable mathematical inferences. Unfortunately, because of outliers, recording, or measurement errors, the assumptions of homoscedasticity are often violated. Ignoring the critical assumptions has an adverse impact on the validity of the test and should be addressed carefully for any version of *t*-test unless the researcher has strong reasons to suppose equal variance (Erceg-Hurn and Mirosevich, 2008). We, therefore, applied Welch's *t*-

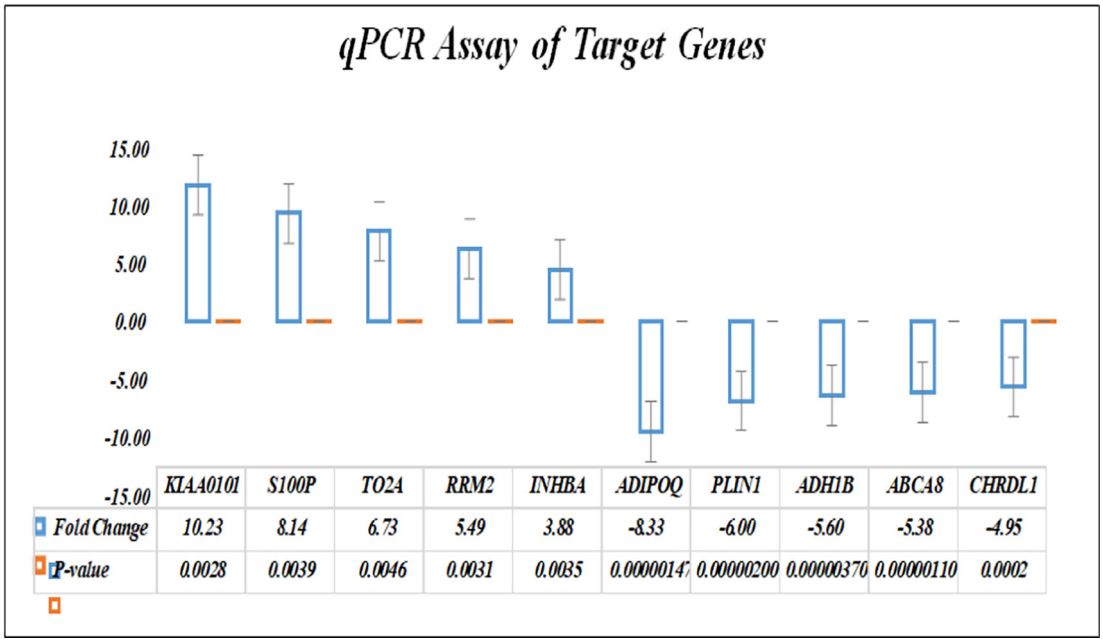

Fig. 3. Bar graph showing quantitative expression of target genes with fold change and *p*-value.

test, a robust statistical method for comparing means to address the assumption of homoscedasticity and generate reliable results (Delacre et al., 2017; Karch, 2021). Next, to evaluate the prognostic values of most significantly genes, we cross-validated them on a survival scale.

Welch's *t-test* identified 1837 DEGs (638 upregulated, 1199 downregulated) which might be playing a vital role in cancer origin and progression. The most significant genes might be a real game changers of breast tumor. However, it was not feasible to discuss the individual role of all genes in one manuscript. We, therefore, are focusing the top 10 up-and down- regulated genes and briefly discussing their diagnostic, prognostic, and therapeutic importance.

Leptin (*LEP*) was the most downregulated gene (FC −35.63), which plays a paramount role in the carcinogenesis of BC (Andò and Catalano, 2011). It increases the proliferation, migration, and invasion of BC cells and could be a novel biomarker for diagnosis, and a potential target for therapeutics (Huang et al., 2017; Maryam et al., 2017). However, Leptin's log rank *p* value was 0.13, more than cut off < 0.05 of KM plot, hence, rejected as a potential prognostic biomarker. Downregulation of adiponectin *C1Q* and collagen domain containing (*ADIPOQ*) (FC −29.72) was reported to be responsible for the primary tumor initiation, maintenance or progression and aggressive BC phenotypes (Mamoor, 2021; Llanos et al., 2020). Perilipin1(*PLIN1*) was downregulated (FC −24.63) in BC as reported earlier and high expression of *PLIN1* indicates longer survival of BC patients (Zhang et al., 2021). *ADIPOQ* and *PLIN1* had log rank *p*-values 0.00071 and 0.0000013 and could be a potential prognostic biomarker.

Phosphoenolpyruvate carboxykinase1 (*PCK1*) and Procollagen C-Endopeptidase Enhancer2 (*PCOLCE2*) were associated with ovarian and BC (Finkernagel et al., 2016). Alcohol dehydrogenase 1B beta polypeptide (*ADH1B*) is well-established cancer biomarker (Polimanti and Gelernter, 2017). *PCK1* and *PCOLCE2* had high log rank *p*-value of 0.06 and 0.84 while *ADH1B* passed the cut-off with a log rank *p* value of 0.00069 for potential to be prognostic biomarkers. Lymphatic Vessel Endothelial Hyaluronan Receptor1 (*LYVE1*) causes disease by altered expression in lymphatic vessel endothelium and used as a cancer marker (Hara et al., 2018). Fatty acid binding protein 4 (*FABP4*) plays a crucial role in tumor progression, particularly in adipose tissue associated cancers by providing fatty acids to the tumor cells (Guaita-Esteruelas et al., 2018). ATP binding cassette subfamily A member8 (*ABCA8*) codes for transporter protein and found significantly downregulated in BC (Hlaváč et al., 2013). Chordin-Like1 (*CHRDL1*) is an established prognostic factor for BC, and downregulation of *CHRDL1* advocates a low survival rate of BC patients (Li et al., 2019). *LYVE1* and *FABP4* did not pass the log rank *p* value cut off, while *ABCA8* (6.30e-11) and *CHRDL1* (4.5e-09) were acceptable as potential prognostic biomarkers in BC.

Collagen type X alpha1 (*COL10A1*) was the most upregulated gene (FC, 22.74) and over expression was reported to enhance the proliferation and metastasis of BC cells (Yang et al., 2020). *KIAA0101* regulates the centrosome of dividing BC cells and enhances cell proliferation and progression (Lv et al., 2018). S100 calcium binding proteinP (*S100P*) increases chemo-resistivity in BC patients and has therapeutic importance (Cong et al., 2020). *COL10A1* with high log rank *p* value was rejected while *KIAA0101* and *S100P* with low log rank *p*-value (1e-^{16} and 6.3e-^{13}) were highly accepted for potential prognostic and therapeutic importance. Overexpression of gap junction protein beta 2 (*GJB2*) is reported in early-stage BC and could be used for an early diagnostic marker (Liu et al., 2019). Topoisomerase II alpha (*TOP2A*) was reported to be linked to tumor grade in early-stage luminal BC (An et al., 2018). Over expression of long intergenic non-protein coding rna 1614 (*LINC01614*) and ribonucleotide reductase M2 (*RRM2*) were associated with overall poor survival of BC patients (Wang et al., 2020; Mazzu et al., 2019). *GJB2* (0.79) and *LINC01614* (0.11) were rejected while *TOP2A* (1.1e-^{16}) and *RRM2* (1e-^{16}) had high acceptance range for potential biomarkers. Over-expression of inhibin betaA (*INHBA*) increases the motility of BC cells (Yu et al., 2021). Chromosome15 open reading frame48 (*C15orf48*) and *Cdc28* protein kinase regulatory subunit2 (*CKS2*) were overexpressed in BC and responsible for initiation and progression (Mamoor, 2021). *INHBA*; *C15orf48* and *CKS2* with log rank *p* values of 0.0017, 0.000046 and 1e-^{16} respectively, were good candidate for prognostic marker. Pedraza et al. focused on the classification of phenotypes with stages of BC, ER (estrogen receptor) status,

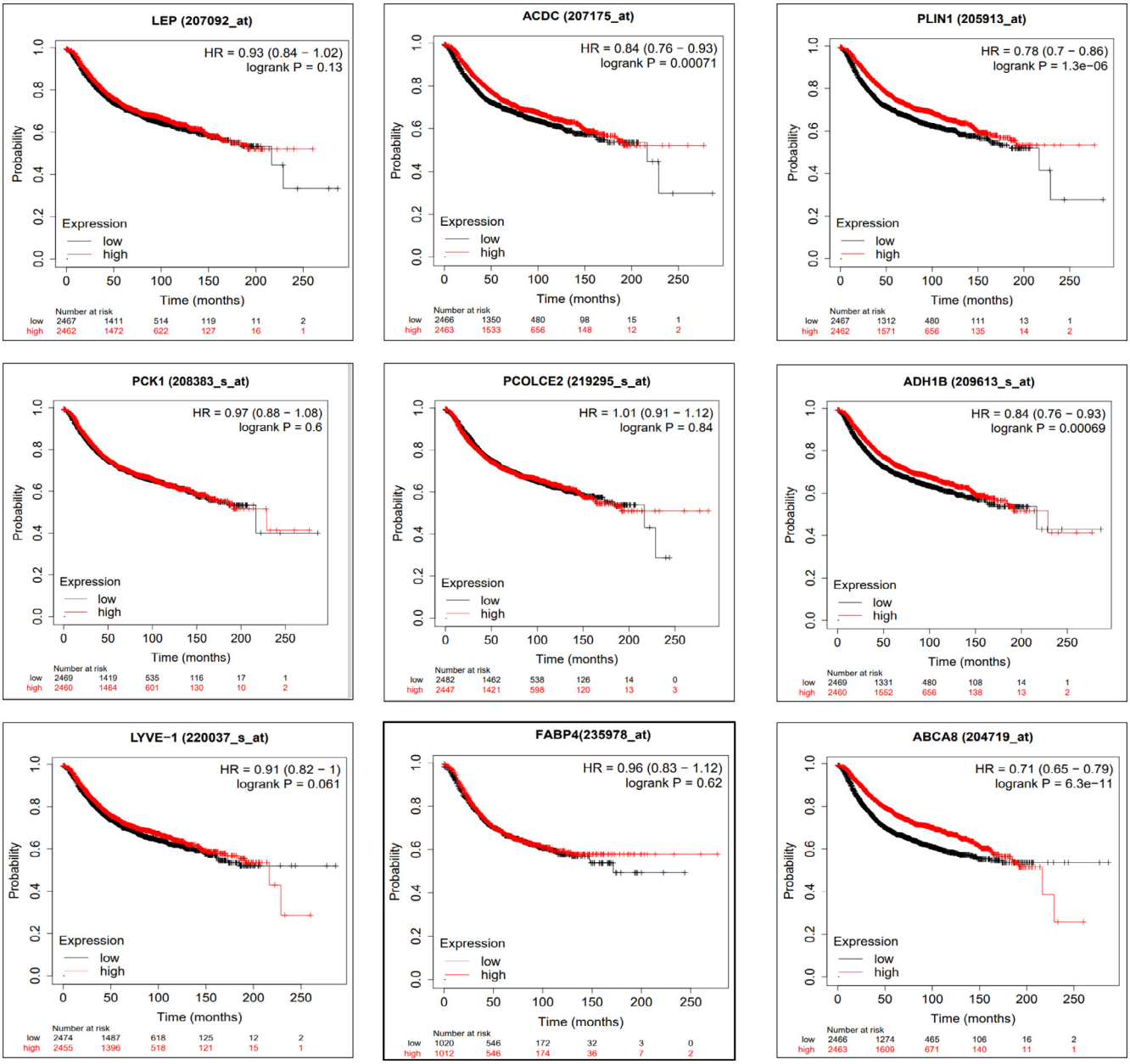

Fig. 4. Figure shows the Kaplan–Meier metastatic relapse-free survival analysis for *LEP, ADIPOQ/ACDC, PLIN1, PCK1, PCOLCE2, LYVE1, FABP4, ABCA8*, and *ADH1B* genes along with the hazard ratio (HR) with 95% confidence intervals (CI) and log rank *p* value.

tumor histology, and lymph node involvement. However, we designed the experiment in a slightly different way wherein focus was concentrated on gene expressions, irrespective of stages of BC, ER status, tumor histology, and lymph node involvement (Kaplan and Meier, 1958).

Additionally, we checked the mathematical relation between BC and control samples via a robust mathematical method (Huber loss robust regression model) and compared it with other regression models (Linear, RANSAC, and Theil-Sen) to get mathematical confidence. Two well-known loss functions are mean square error (MSE, L2 Loss) and mean absolute error (MAE, L1 Loss), and both have some advantages and disadvantages. We applied a mixture of MSE/L2 and MAE/L1 in Huber loss robust regression model as it is sensitive to outliers than the squared error loss, differentiable at zero, the error is squared for small values. It gives less weight to outliers with extreme values. Based on the solid theoretical and mathematical justification, the result showed a weak relationship between tumor and control samples, as both groups are different from each other.

Previous group had used a moderate student *t* test for normal data and Mann Whitney test for non-normal data analysis (Kaplan and Meier, 1958). However, we have gone one step further as a complementary approach with underlying assumptions of the mathematical model using robust Welch's *t* test and tried to correlate control and tumors samples through the Huber loss robust regression model. Additionally, first, we cross-validated the reanalysis results using Kaplan– Meier plot to examine if the expression values are linked with the overall metastatic relapse- free survival of BC patients and second confirmed the expression level by qPCR assay on bigger cohort of BC. Thus the significantly expressed genes have prognosis, diagnosis, and therapeutics potential and needs to be further evaluated.

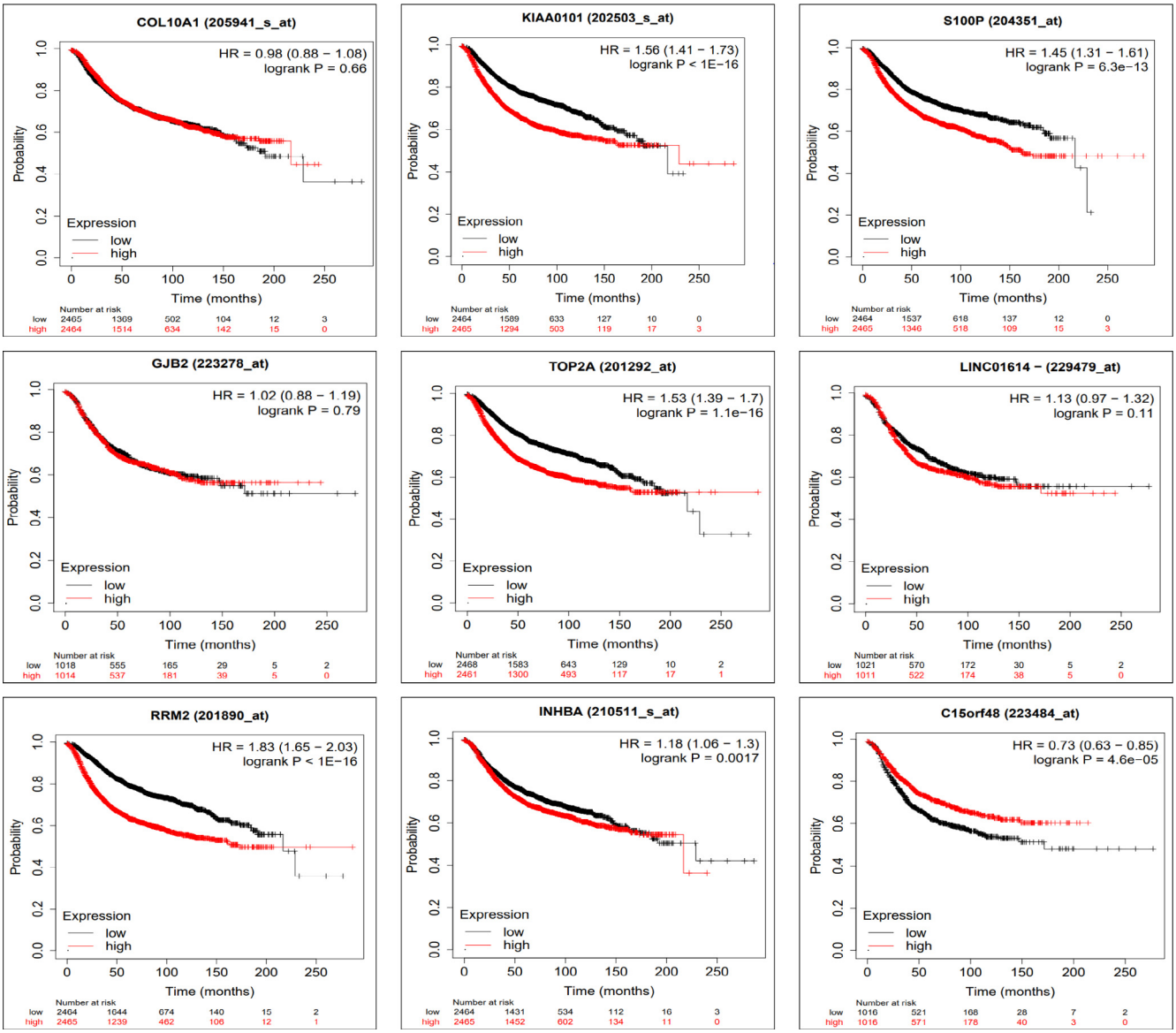

Fig. 5. Figure shows the Kaplan–Meier metastatic relapse-free survival analysis for *COL10A1*, *KIAA0101*, *S100P*, *GJB2*, *TOP2A*, *RRM2*, *INHBA*, *C15orf48*, and *LINC01614* genes along with the hazard ratio (HR) with 95% confidence intervals (CI) and log rank *p* value.

Table 2
Kaplan–Meier Plot values for the top 10 up- and down-expressed genes.

Gene Symbol	Fold Change	Hazard Ratio (HR)	Confidence Interval (95 %)	Log rank *p* value	Decision
COL10A1	22.74	0.98	0.88–1.08	0.66	Reject
KIAA0101	12.61	1.56	1.41–1.73	< 1e-16	Accept
S100P	11.54	1.45	1.31–1.61	6.3E-13	Accept
GJB2	11.03	1.02	0.88–1.19	0.79	Reject
TOP2A	10.43	1.53	1.39–1.70	1.1E-16	Accept
LINC01614	9.99	1.13	0.97–1.32	0.11	Reject
RRM2	9.68	1.83	1.65–2.03	< 1e-16	Accept
INHBA	8.70	1.18	1.06–1.30	0.0017	Accept
C15orf48	8.56	0.73	0.63–0.85	4.6E-05	Accept
CKS2	7.72	1.67	1.51–1.85	< 1e-16	Accept
LEP	−35.64	0.93	0.84–1.02	0.13	Reject
ADIPOQ	−29.73	0.84	0.76–0.93	0.00071	Accept
PLIN1	−24.63	0.78	0.70–0.86	1.30E-06	Accept
PCK1	−23.93	0.97	0.88–1.08	0.06	Reject
PCOLCE2	−22.83	1.01	0.91–1.12	0.84	Reject
ADH1B	−21.54	0.84	0.76–0.93	0.00069	Accept
LYVE1	−20.20	0.91	0.82–1.00	0.061	Reject
FABP4	−20.12	0.96	0.83–1.12	0.062	Reject
ABCA8	−19.42	0.71	0.65–0.79	6.30E-11	Accept
CHRDL1	−18.49	0.74	0.67–0.82	4.50E-09	Accept

Table 3
Comparison of model performance based on mean absolute error, standard deviation, regression coefficient, regression intercept and Equation of best Fit Line for Linear, Huber, RANSAC and TheilSen regression.

Models/Algorithms	MAE	Standard Deviation	Regression Coefficient	Regression Intercept	Mean Tumor = coefficient × Mean Control + intercept
Linear Regression	1.075	0.02	0.50353571	3.09092	0.5035* Mean Control + 3.0909,
Huber Regression	1.069	0.02	0.47639532	3.219997	0.4763* Mean Control + 3.2199
RANSAC Regression	1.245	0.105	0.614	1.676	0.6140* Mean Control + 1.6700
TheilSen Regression	1.093	0.018	0.58977372	2.471666	0.5897* Mean Control + 2.4716

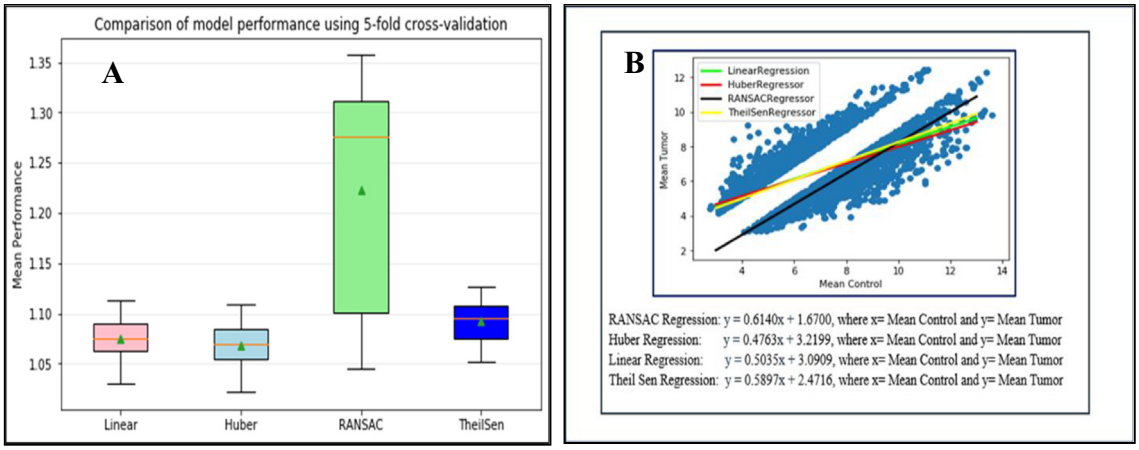

Fig. 6. **(A)** Box and Whisker plot for Linear, Huber, RANSAC, and Theil-Sen regression models and **(B)**: Scattered diagram with the best fit line through Linear, Huber, RANSAC and TheilSen regression models for comparative 3126 probes mean gene expression values in normal and tumor samples.

5. Conclusions

Statistical method like Welch Satterthwaite *t*-test and Huber loss robust regression model algorithms gave mathematical confidence in detecting DEGs and improved the understanding of microarray gene expression profiling of BC. It revealed a weak mathematical relation (co– relation coefficient: 0.43) that represents the differences between tumor and control samples. Using minimum underlying assumptions for Welch Satterthwaite *t*-test and Kaplan-Meier estimator plot models were novel approach. Refined survival analysis of most significantly expressed genes showed twelve genes correlated with the overall metastatic relapse-free survival. Finally, ten clinically associated genes were validated by qPCR that may be promising diagnosis, prognosis, and/or therapeutics biomarkers of BC.

CRediT authorship contribution statement

Sajjad Karim: Conceptualization, Supervision, Project administration, Writing – original draft. **Md Shahid Iqbal:** Conceptualization, Data curation, Investigation, Methodology, Writing – original draft. **Nesar Ahmad:** Conceptualization, Supervision. **Md Shahid Ansari:** Formal analysis, Data curation, Writing – review & editing. **Zeenat Mirza:** Methodology, Writing – original draft. **Adnan Merdad:** Data curation, Writing – review & editing. **Saddig D. Jastaniah:** Project administration. **Sudhir Kumar:** Conceptualization, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was funded by King Abdulaziz University under grant No. (2-117-1434-HiCi). The authors acknowledge the technical and financial support of KAU. We would also like to thank AZIZ Supercomputing facilities at High Performance Computing Center for their help and support.

Disclosure of funding

This paper was funded by King Abdulaziz University, under grant No. (2-117-1434-HiCi). The authors, therefore, acknowledge technical and financial support of KAU.

Ethics Approval and Consents:

Ethic approval and consent of participants do not apply here as the study was based on statistical analysis of data retrieved from public database (GEO, NCBI). However, biobank samples were used for validation study.

Availability of Data

The clinicopathological information and datasets (.CEL file) supporting the results of this article were submitted to NCBI's Gene Expression Omnibus (GEO) under accession number GSE10810.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jksus.2022.102447.

References

An, X., Xu, F., Luo, R., Zheng, Q., Lu, J., Yang, Y., Qin, T., Yuan, Z., Shi, Y., Jiang, W., Wang, S., 2018. The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. BMC Cancer 18 (1), pp. https://doi.org/10.1186/s12885-018-4170-7.

Andò, S., Catalano, S., 2011. The multifactorial role of leptin in driving the breast cancer microenvironment. Nat. Rev. Endocrinol. 8 (5), 263–275. https://doi.org/10.1038/nrendo.2011.184.

Baguley, T. 2012. "A guide to advanced statistics for the behavioral sciences," Serious stats. Houndmills: Palgrave Macmillan, 2012, [Online]. Available: https://books.google.fr/books?hl=fr&lr=&id=ObUcBQAAQBAJ&oi=fnd&pg=PP1&dq=baguley+2012 &ots=-eiUlHiCYs&sig= YUUKZ7jiGF33wdo3 WVO-8l-OUu8.

"Breast cancer", *Who.int*, 2021, [Online]. Available: https://www.who.int/news-room/fact- sheets/detail/breast-cancer. [Accessed: 10- Sep- 2021].

Cong, Y., Cui, Y., Wang, S., Jiang, L., Cao, J., Zhu, S., Birkin, E., Lane, J., Ruge, F., Jiang, W., Qiao, G., 2020. Calcium-Binding Protein S100P Promotes Tumor Progression but Enhances Chemosensitivity in Breast Cancer. Front. Oncol. 10. https://doi.org/10.3389/fonc.2020.566302.

Cox, D., 1972. Regression Models and Life-Tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) 34 (2), 187–202.

de Siqueira Santos, S., Takahashi, D., Nakata, A., Fujita, A., 2013. A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief. Bioinform. 15 (6), 906–918. https://doi.org/10.1093/bib/bbt051.

Delacre, M., Lakens, D., Leys, C., 2017. Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. International Review of Social Psychology 30 (1), 92. https://doi.org/10.5334/irsp.82.

Emmert-Streib, F., Dehmer, M., 2019. Introduction to Survival Analysis in Practice. Machine Learning and Knowledge Extraction 1 (3), 1013–1038. https://doi.org/10.3390/make1030058.

Erceg-Hurn, D., Mirosevich, V., 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. Am. Psychol. 63 (7), 591–601. https://doi.org/10.1037/0003-066x.63.7.591.

Finkernagel, F., Reinartz, S., Lieber, S., Adhikary, T., Wortmann, A., Hoffmann, N., Bieringer, T., Nist, A., Stiewe, T., Jansen, J., Wagner, U., Müller-Brüsselbach, S., Müller, R., 2016. The transcriptional signature of human ovarian carcinoma macrophages is associated with extracellular matrix reorganization. Oncotarget 7 (46), 75339–75352. https://doi.org/10.18632/oncotarget.12180.

Guaita-Esteruelas, S., Gumà, J., Masana, L., Borràs, J., 2018. The peritumoural adipose tissue microenvironment and cancer. The roles of fatty acid binding protein 4 and fatty acid binding protein 5. Mol. Cell. Endocrinol. 462, 107–118. https://doi.org/10.1016/j.mce.2017.02.002.

Gupta, A., Mishra, P., Pandey, C., Singh, U., Sahu, C., Keshri, A., 2019. Descriptive statistics and normality tests for statistical data. Ann. Card. Anaesth. 22 (1), 67. https://doi.org/10.4103/aca.aca_157_18.

Hara, Y., Torii, R., Ueda, S., Kurimoto, E., Ueda, E., Okura, H., Tatano, Y., Yagi, H., Ohno, Y., Tanaka, T., Masuko, K., Masuko, T., 2018. Inhibition of tumor formation and metastasis by a monoclonal antibody against lymphatic vessel endothelial hyaluronan receptor 1. Cancer Sci. 109 (10), 3171–3182. https://doi.org/10.1111/cas.13755.

Hlaváč, V., Brynychová, V., Václavíková, R., Ehrlichová, M., Vrána, D., Pecha, V., Koževnikovová, R., Trnková, M., Gatěk, J., Kopperová, D., Gut, I., Souček, P., 2013. The expression profile of ATP-binding cassette transporter genes in breast carcinoma. Pharmacogenomics 14 (5), 515–529. https://doi.org/10.2217/pgs.13.26.

Huang, Y., Jin, Q., Su, M., Ji, F., Wang, N., Zhong, C., Jiang, Y., Liu, Y., Zhang, Z., Yang, J., Wei, L., Chen, T., Li, B., 2017. Leptin promotes the migration and invasion of breast cancer cells by upregulating ACAT2. Cell. Oncol. 40 (6), 537–547. https://doi.org/10.1007/s13402-017-0342-8.

Kaplan, E., Meier, P., 1958. Nonparametric Estimation from Incomplete Observations. J. Am. Stat. Assoc. 53 (282), 457–481.

Karch, J. 2021. "Choosing Between the Two-sample T Test and its Alternatives: A Practical Guideline," doi: 10.31234/osf.io/ye2d4.

Lánczky, A., Győrffy, B., 2021. Web-Based Survival Analysis Tool Tailored for Medical Research (KMplot): Development and Implementation. J. Med. Internet Res. 23 (7), e27633.

Li, D., Li, L., Cao, Y., Chen, X., 2019. Downregulation of LINC01140 is associated with adverse features of breast cancer. Oncol. Lett. https://doi.org/10.3892/ol.2019.11147.

Liu, Y., Pandey, P., Sharma, S., Xing, F., Wu, K., Chittiboyina, A., Wu, S., Tyagi, A., Watabe, K., 2019. ID2 and GJB2 promote early-stage breast cancer progression by regulating cancer stemness. Breast Cancer Res. Treat. 175 (1), 77–90. https://doi.org/10.1007/s10549-018-05126-3.

Llanos, A., Yao, S., Singh, A., Aremu, J., Khiabanian, H., Lin, Y., Omene, C., Omilian, A., Khoury, T., Hong, C., Ganesan, S., Foran, D., Higgins, M., Ambrosone, C., Bandera, E., Demissie, K., 2020. Gene expression of adipokines and adipokine receptors in the tumor microenvironment: associations of lower expression with more aggressive breast tumor features. Breast Cancer Res. Treat. 185 (3), 785–798. https://doi.org/10.1007/s10549-020-05972-0.

Lv, W., Su, B., Li, Y., Geng, C., Chen, N., 2018. KIAA0101 inhibition suppresses cell proliferation and cell cycle progression by promoting the interaction between p53 and Sp1 in breast cancer. Biochem. Biophys. Res. Commun. 503 (2), 600–606. https://doi.org/10.1016/j.bbrc.2018.06.046.

Mamoor, S. 2021. "Differential expression of CKS2 in cancers of the breast," 2021, doi: 10.31219/osf.io/zdhr3.

Mamoor, S. 2021. "Differential expression of adiponectin, C1Q and collagen domain containing in cancers of the breast," 2021, doi:10.31219/osf.io/9zcx8.

Maryam, R., Faegheh, S., Majid, A., Kazem, N., 2017. Effect of quercetin on secretion and gene expression of leptin in breast cancer. J. Tradit. Chin. Med. 37 (3), 321–325. https://doi.org/10.1016/s0254-6272(17)30067-5.

Mazzu, Y., Armenia, J., Chakraborty, G., Yoshikawa, Y., Coggins, S., Nandakumar, S., Gerke, T., Pomerantz, M., Qiu, X., Zhao, H., Atiq, M., Khan, N., Komura, K., Lee, G., Fine, S., Bell, C., O'Connor, E., Long, H., Freedman, M., Kim, B., Kantoff, P., 2019. A Novel Mechanism Driving Poor-Prognosis Prostate Cancer: Overexpression of the DNA Repair Gene, Ribonucleotide Reductase Small Subunit M2 (RRM2). Clin. Cancer Res. 25 (14), 4480–4492. https://doi.org/10.1158/1078-0432.

Parkin, D., 2006. The global health burden of infection-associated cancers in the year 2002. Int. J. Cancer 118 (12), 3030–3044. https://doi.org/10.1002/ijc.21731.

Pedraza, V., Gomez-Capilla, J.A., Escaramis, G., Gomez, C., Torné, P., Rivera, J.M., Gil, A., Araque, P., Olea, N., Estivill, X., Fárez-Vidal, M.E., 2010. Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. Cancer 116 (2), 486–496.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., Franceschi, S., 2016. Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob. Health 4 (9), e609–e616. https://doi.org/10.1016/s2214-109x(16)30143-7.

Polimanti, R., Gelernter, J., 2017. ADH1B: From alcoholism, natural selection, and cancer to the human phenome. American Journal Of Medical Genetics Part B: Neuropsychiatric Genetics 177 (2), 113–125. https://doi.org/10.1002/ajmg.b.32523.

Rasch, D., Kubinger, K., Moder, K., 2009. The two-sample t test: pre-testing its assumptions does not pay off. Stat. Pap. 52 (1), 219–231. https://doi.org/10.1007/s00362-009-0224-x.

Russo, G., Zegar, C., Giordano, A., 2003. Advantages and limitations of microarray technology in human cancer. Oncogene 22 (42), 6497–6507. https://doi.org/10.1038/sj.onc.1206865.

Wang, D., Zhang, H., Fang, X., Cao, D., Liu, H., 2020. Pan-cancer analysis reveals the role of long non-coding RNA LINC01614 as a highly cancer-dependent oncogene and biomarker. Oncol. Lett. 20 (2), 1383–1399. https://doi.org/10.3892/ol.2020.11648.

Yang, W., Wu, X., Zhou, F., 2020. Collagen Type X Alpha 1 (COL10A1) Contributes to Cell Proliferation, Migration, and Invasion by Targeting Prolyl 4-Hydroxylase Beta Polypeptide (P4HB) in Breast Cancer. Med. Sci. Monit. 27. https://doi.org/10.12659/MSM.928919.

Yu, Y., Wang, W., Lu, W., Chen, W., Shang, A., 2021. Inhibin β-A (INHBA) induces epithelial– mesenchymal transition and accelerates the motility of breast cancer cells by activating the TGF-β signaling pathway. Bioengineered 12 (1), 4681–4696. https://doi.org/10.1080/21655979.2021.1957754.

Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., Moch, H., Cooper, C.S., Eils, R., Ferretti, V., Lichter, P., 2020. The landscape of viral associations in human cancers. Nat. Genet. 52 (3), 320–330.

Zhang, X., Su, L., Sun, K., 2021. Expression status and prognostic value of the perilipin family of genes in breast cancer [Online].Available: Am. J. Transl. Res. 13 (5), 4450–4463 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8205812/.