JKSUS revision

By shahid

WORD COUNT 3811
TIME SUBMITTED 15-SEP-2022 11:34AM
PAPER ID 90537117

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

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 has high potential to be BC biomarkers using Welch t-test and Kaplan-Meier plot having minimum underlying assumptions.

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

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 [1-3]. 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 [4]. 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 [5]. 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 [6]. 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 [7]. 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 [8-9].

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 [10-11].

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.

Materials and Methods

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 54675 probes [12]. The cohort contains 58 samples including 31 BC and 27 control.

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:

$$\omega(t)=(\Delta\bar{X})/s_{\Delta\bar{X}} = (\bar{X}_1-\bar{X}_2)/\sqrt{(s_{\bar{X}_1})^2+s_{(\bar{X}_2)^2}} \quad (1)$$

$$s_{(\bar{X}_i)}=s_i/\sqrt{N_i} \quad (2)$$

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 = ω(v)combined with this variance estimate,

is approximated using the Welch–Satterthwaite equation

$$\omega(v)\approx((s_1^2)/N_1 +(s_2^2)/N_2)^2/((s_1^4)/(N_1^2\ \omega(v_1))+(s_2^4)/(N_2^2\ \omega\ 〚(v〛_2))) \quad (3)$$

In case of N1=N2

$$\omega(v) \approx \langle\!\langle s^4 \rangle\!\rangle_{(\Delta \bar{X})} / (\omega(v_1)^{-1} \langle\!\langle s^4 \rangle\!\rangle_{(\bar{X}_1)} + \omega(v_2)^{-1} \langle\!\langle s^4 \rangle\!\rangle_{(\bar{X}_2)})$$

(4)

$\langle\!\langle \omega(v) \rangle\!\rangle_i = N_i - 1$ 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 [13]. Based on cut-off p-values ≤ 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.

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 [10] is mathematically expressed as:

$$\text{SF at } t = \prod_{(i=t_i<t)} (n_i - d_i)/n_i = \prod_{(i=t_i<t)} (1 - d_i/n_i) \quad (5)$$

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 and t_i+1. Equation 5 can be rewritten using a recursive formula.

SF at t_j = [(n_(j-1)-d_(j-1))/n_(j-1)] multiply by (SF) at t_(j-2)

(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% confidence interval, calculate hazard ratio (HR), statistical significance log rank p value was ≤ 0.05 (https://kmplot.com/analysis/) [14-15].

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 [16]. 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:

$$MSE=(\sum_{i=1}^{n}(y_i-y_i^P)^2)/n \quad (7)$$

$$MAE=(\sum_{i=1}^{n}(|y_i-\llbracket y_i \rrbracket^p|))/n \quad (8)$$

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 (□, delta).

$$L_\delta(y,f(x))=\begin{cases}(1/2)(y-\llbracket f(x) \rrbracket)^2 & \text{for } |y-f(x)|\le\delta, \\ @\delta|y-f(x)|-1/2\,\delta^2 & \text{otherwise}\end{cases} \quad (9)$$

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.

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.

Results

A total of 54675 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 (7.414±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 probe 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 54675 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 (Figure 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 54675 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

Figure 1: PP (Probability-Probability) and QQ (Quantile-Quantile) plots for the tumor and control variables with 54675 probes mean gene expression values for normality

check as the underlying assumption of Welch t-test: (A) PP plot of tumor, (B) PP plot of control, (C) QQ plot of tumor, and (D) QQ plot of control.

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 (Figure 2).

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

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 (Figure 3)

	KIAA0101	S100P	TO2A	RRM2	INHBA	ADIPOQ	PLIN1	ADH1B	ABCA8	CHRDL1
Fold Change	10.23	8.14	6.73	5.49	3.88	-8.33	-6.00	-5.60	-5.38	-4.95
P-value	0.0028	0.0039	0.0046	0.0031	0.0035	0.00000147	0.00000200	0.00000370	0.00000110	0.0002

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

Kaplan-Meier Plotter was used to confirm survival in a larger dataset and its association with the identified genes. KM Plot (Figure 4-5) 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.

Figure 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.

Figure 5: Figure shows the Kaplan–Meier metastatic relapse-free survival analysis for COL10A1, KIAA0101, S100P, GJB2, TOP2A, RRM2, INHBA, C15orf48, CKS2 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

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 was 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, Figure 6).

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 Huber	1.075	0.02	0.50353571	3.09092	0.5035* Mean Control + 3.0909,
Regression RANSAC	1.069	0.02	0.47639532	3.219997	0.4763* Mean Control + 3.2199
Regression TheilSen	1.245	0.105	0.614	1.676	0.6140* Mean Control + 1.6700
Regression	1.093	0.018	0.58977372	2.471666	0.5897* Mean Control + 2.4716

Figure 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.

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 unbiased, 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 [7]. We, therefore, applied Welch's t-test, a robust statistical method for comparing means to address the assumption of homoscedasticity and generate reliable results [9], [17]. 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 [18]. It increases the proliferation, migration, and invasion of BC cells and could be a novel biomarker for diagnosis, and a potential target for therapeutics [19-20]. 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 [21- 22]. Perilipin1(PLIN1) was

downregulated (FC -24.63) in BC as reported earlier and high expression of PLIN1 indicates longer survival of BC patients [23]. 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 [24]. Alcohol dehydrogenase 1B beta polypeptide (ADH1B) is well-established cancer biomarker [25]. 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 [26]. 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 [27]. ATP binding cassette subfamily A member8 (ABCA8) codes for transporter protein and found significantly downregulated in BC [28]. Chordin-Like1 (CHRDL1) is an established prognostic factor for BC, and

downregulation of CHRDL1 advocates a low survival rate of BC patients [29]. 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 [30]. KIAA0101 regulates the centrosome of dividing BC cells and enhances cell proliferation and progression [31]. S100 calcium binding proteinP (S100P) increases

chemo-resistivity in BC patients and has therapeutic importance [32]. 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 [33]. Topoisomerase II alpha (TOP2A) was reported to be linked to tumor grade in early-stage luminal BC [34]. Over expression of long intergenic non-protein coding rna 1614 (LINC01614) and ribonucleotide reductase M2 (RRM2) was associated with overall poor survival of BC patients [35-36]. GJB2 (0.79) and LINC01614 (0.11) was 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 [37]. Chromosome15 open reading frame48 (C15orf48) and Cdc28 protein kinase regulatory subunit2 (CKS2) were overexpressed in BC and responsible for initiation and progression [38]. 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, 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 [10].

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 [10]. However, we have gone one step further as a complementary approach underlying assumptions of the mathematical model using robust Welch's 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.

Conclusions

Statistical method like Welch Satterthwaite t-test and Huber loss robust regression model algorithms gave mathematical confidence in detecting DEGs and improve 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 was noble 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.

JKSUS revision

ORIGINALITY REPORT

13%
SIMILARITY INDEX

PRIMARY SOURCES

1	heartbeat.fritz.ai Internet	41 words — 1%
2	www.rips-irsp.com Internet	41 words — 1%
3	Oluwatosin E. Bankole, Deepak Kumar Verma, Mónica L. Chávez González, Juan Guzmán Ceferino et al. "Recent trends and technical advancements in biosensors and their emerging applications in food and bioscience", Food Bioscience, 2022 Crossref	30 words — 1%
4	portlandpress.com Internet	30 words — 1%
5	www.researchgate.net Internet	29 words — 1%
6	en.wikipedia.org Internet	22 words — 1%
7	Marie Delacre, Daniël Lakens, Christophe Leys. "Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test", International Review of Social Psychology, 2017 Crossref	21 words — < 1%

8	www.ncbi.nlm.nih.gov Internet	21 words — < 1%
9	R.J. Kauffman, Bin Wang. "Duration in the digital economy", 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the, 2003 Crossref	19 words — < 1%
10	www.mdpi.com Internet	19 words — < 1%
11	Oseni, Saheed Oluwasina. "Role of Interleukin-1 Receptor-Associated Kinases in Chronic Inflammation and Prostate Tumorigenesis", Florida Atlantic University, 2021 ProQuest	15 words — < 1%
12	shuang wu, Shihai Liu, Yongxian Cao, Chao Geng, Peng Wang, Huazheng Pan. "Downregulated ZC3H13 by miR-362-3p/miR-425-5p is Associated with Poor Prognosis and Adverse Outcomes in Hepatocellular Carcinoma", Research Square Platform LLC, 2021 Crossref Posted Content	15 words — < 1%
13	www.medrxiv.org Internet	15 words — < 1%
14	Research-Repository.griffith.edu.au Internet	14 words — < 1%
15	tuprints.ulb.tu-darmstadt.de Internet	14 words — < 1%
16	vtechworks.lib.vt.edu Internet	14 words — < 1%
17	scholarcommons.usf.edu	

	Internet	13 words — < 1%
18	WWW.MDPI.COM Internet	12 words — < 1%
19	dokumen.pub Internet	12 words — < 1%
20	pubmed.ncbi.nlm.nih.gov Internet	11 words — < 1%
21	www.frontiersin.org Internet	11 words — < 1%
22	Bengt Gunnarsson. "Maintenance of melanism in the spider Pityohyphantes phrygianus: is bird predation a selective agent?", Heredity, 1993 Crossref	10 words — < 1%
23	Gabriella Iván. "Survey of free speeds on rural roads based on road scene photographs", Pollack Periodica, 2012 Crossref	10 words — < 1%
24	academic.oup.com Internet	10 words — < 1%
25	acrabstracts.org Internet	10 words — < 1%
26	bmcgenomics.biomedcentral.com Internet	10 words — < 1%
27	jneuroinflammation.biomedcentral.com Internet	10 words — < 1%

28	Barreiro-Arevalo, Myrine A.. "Significant Gene Array Analysis and Cluster-Based Machine Learning for Disease Class Prediction", The University of Texas Rio Grande Valley, 2021 ProQuest	9 words — < 1%
29	mdpi-res.com Internet	9 words — < 1%
30	preview-almob.biomedcentral.com Internet	9 words — < 1%
31	AtlasGeneticsOncology.org Internet	8 words — < 1%
32	Wrenn, Emma. "Tumor Cell Clusters Form Intercellular Nanolumina Which Promote Collective Metastasis Via Epigen Accumulation", University of Washington, 2021 ProQuest	8 words — < 1%
33	roderic.uv.es Internet	8 words — < 1%
34	www.ijmp.jor.br Internet	8 words — < 1%
35	René L. Schilling. "13 Strassen's functional law of the iterated logarithm", Walter de Gruyter GmbH, 2021 Crossref	7 words — < 1%
36	digitalcommons.liberty.edu Internet	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF