AB Journals

African Journal of Mathematics and Statistics Studies
ISSN 2689-5323 • Volume 9, Issue 3, 2026 • pp. 45–61

RESEARCH ARTICLE Received: 6 Jun 2026, Accepted: 5 Jul 2026, Published: 11 Sep 2026

Analyzing Gene Expression Patterns in Breast Cancer Subtypes Using Multivariate Analysis of Variance (MANOVA)

Thomas Adidaumbe Ugbe*; **Nku George Ekong**

Department of Statistics, Faculty of Physical Sciences, University of Calabar.

**Corresponding Author's Email: ugbethomas1@gmail.com*

Abstract

Aims: This study investigated the expression patterns of five crucial genes BRCA1, PALB2, CDK1, CDH1, and CHEK2 across breast cancer subtypes using Multivariate Analysis of Variance (MANOVA).

Study Design: A quantitative research design based on MANOVA was employed using gene expression data from 1,139 breast cancer patients representing multiple subtypes.

Place and Duration of Study: The study utilized secondary genomic data obtained from publicly available breast cancer databases and was conducted over a six-month period.

Methodology: Preliminary diagnostic tests assessed normality, multicollinearity, and multivariate outliers. Shapiro–Wilk tests indicated approximate normality, VIF values below 5 confirmed the absence of multicollinearity, and Mahalanobis distance showed no significant outliers ($12.68 < 20.515$). Although Levene's and Box's M tests indicated heterogeneity of variances and covariance matrices ($p < 0.001$), Pillai's trace was adopted due to its robustness. Univariate analyses and Scheffe post hoc tests were subsequently conducted.

Results: MANOVA revealed a highly significant overall effect of breast cancer subtype on combined gene expression profiles (Pillai's Trace = 0.525, $F = 48.09$, $p < 0.001$, observed power = 1.000). BRCA1, CDK1, CDH1, and CHEK2 significantly discriminated subtypes ($p < 0.001$), while PALB2 showed marginal significance ($p = 0.003$). Luminal A and HER2 demonstrated the strongest gene-specific effects, whereas Luminal B showed the least differentiation.

Conclusion: The findings confirm distinct subtype-specific gene expression patterns and demonstrate the effectiveness of MANOVA for genomic studies, highlighting its relevance for precision oncology and targeted breast cancer management.

Keywords: Breast cancer, Gene expression, Genomics, Multicollinearity, Subtypes.

How to Cite: Thomas, A. U., & Nku, G. E. (2026). Analyzing Gene Expression Patterns in Breast Cancer Subtypes Using Multivariate Analysis of Variance (MANOVA). African Journal of Mathematics and Statistics Studies, 9(3), 45–61. https://doi.org/10.52589/AJMSS-CUY8H5GT

INTRODUCTION

Breast cancer, which accounts for over 30% of all new cancer diagnoses in women and causes over 685,000 deaths annually worldwide, is a diverse illness that poses a serious threat to global health (Sung et al., 2021). The molecular categorization of breast cancer into four distinct subtypes luminal A, luminal B, HER2-enriched, and triple-negative breast cancer has transformed methods for diagnosis, prognosis, and treatment. The expression patterns of three important receptors, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), are largely responsible for defining these subtypes. Among the subtypes of breast cancer, luminal A cancers have the best prognosis, strong ER/PR expression, low proliferation rates, and frequently respond effectively to endocrine treatments (Yin et al., 2020). On the other hand, luminal B tumors are more aggressive and require combined endocrine and chemotherapy treatments due to their increased proliferation (e.g., enhanced Ki-67) and potential HER2 positivity (Alqahtani et al., 2025). Targeted anti-HER2 treatments like trastuzumab have successfully treated HER2-enriched subtypes, which are characterized by HER2 amplification without ER/PR expression and were previously linked to poor outcomes (Yershova & Singhal, 2022). TNBC, which accounts for 15–20% of instances of breast cancer and lacks ER, PR, and HER2 expression, is infamous for its aggressiveness, high recurrence rates, and lack of tailored treatment choices, which frequently depend on chemotherapy (Wang et al., 2025).

By incorporating gene expression profiles, recent developments in breast cancer molecular subtyping, such as the PAM50 assay, have further refined these categories and revealed additional heterogeneity within subtypes, such as TNBC's division into basal-like, mesenchymal, and immunomodulatory groups (Chen et al., 2025).

Multivariate analysis of variance (MANOVA) is used in this work to examine the expression patterns of important genes across molecular subtypes of breast cancer. The goal of the investigation is to find coordinated variations that guide treatment and prognosis. It fills knowledge gaps in subtype heterogeneity by incorporating new genetic insights, which may help direct precision therapy.

An expansion of univariate ANOVA, multivariate analysis of variance (MANOVA) is used to examine situations with several dependent variables at once. Wilks (1932) and Hotelling (1931) laid the groundwork for MANOVA, which offers a coherent framework for analyzing intricate, multidimensional data sets. Since genes frequently have coordinated expression patterns that imply similar regulatory mechanisms or functional pathways, this technique is very useful in genomic research (Langfelder & Horvath, 2008; Stevens, 2009). When several genes are analyzed collectively, MANOVA can find minor but significant patterns that may go unnoticed when genes are analyzed separately (Huberty & Olejnik, 2006).

Particularly in genomic applications, MANOVA functions under a number of fundamental assumptions that need to be carefully considered. All dependent variables and their linear combinations must have a normal distribution within each group in order to satisfy the multivariate normality condition. Although skewness or outliers frequently cause gene expression data to defy this premise, these problems can be lessened by using transformation methods such as variance-stabilizing transformations or log-transformation (Law et al., 2014; Tabachnick & Fidell, 2013). The homogeneity of variance-covariance matrices, which states that the correlation structure among dependent variables should be comparable across all groups, is another crucial premise. Although there are strong alternatives, violations of this premise can be troublesome, especially in small-sample studies (Olson, 1976; Warne, 2014).

MANOVA provides several test statistics that are appropriate for various research situations. The most widely used statistic, Wilks' Lambda, shows the ratio of within-group to total variance and performs well overall under a variety of circumstances (Wilks, 1932). Because Pillai's Trace is resistant to assumption violations and unequal sample sizes, it is especially well-suited for genomic investigations, where these circumstances are common (Olson, 1976). Roy's Largest Root is strong but susceptible to assumption violations, whereas Hotelling's Trace is best for comparing just two groups (Huberty & Petoskey, 2000). Large-scale genomic investigations have improved the intrinsic molecular subgroups of breast cancer, which were initially identified by Perou et al. (2000).

Excellent 5-year survival rates are provided by luminal A, which accounts for 50–60% of cases and is ER/PR-positive, HER2-negative, and has low Ki-67 (Laenkholm et al., 2025). With an intermediate prognosis of 70–80% 5-year survival and resistance to endocrine therapy, Luminal B, which accounts for 20–30% of patients, shares ER/PR positivity but shows increased proliferation and sporadic HER2 amplification (Guo et al., 2025). Targeted therapies have improved outcomes for HER2-enriched tumors (10–15%), which are aggressive and develop quickly due to HER2 overexpression (Alqahtani et al., 2025). With diverse subtypes such as basal-like (DNA repair deficient), mesenchymal (EMT-driven), and immunomodulatory (immune-infiltrated), TNBC (15–20%) lacks targeted options (Alghamdi et al., 2025). TNBC's promise for immunotherapy is highlighted by recent classifications that include immunological scores (Chen et al., 2025).

Different breast cancer subtypes have different clinical implications. Hormone therapy is beneficial for luminal subtypes, monoclonal antibodies for HER2-enriched cancers, and PARP inhibitors for TNBC in instances with BRCA mutations (Rossing et al., 2021). According to prevalence studies, younger patients are more likely to have luminal B (Santonja et al., 2024). A key mechanism in molecular biology, gene expression acts as a link between genotype and phenotype. It describes the method by which functional gene products, usually proteins, are created using information from a gene (Hanahan & Weinberg, 2011). The quantity of these products generated in a cell or tissue at a certain moment is represented by the level of gene expression, which is important for cell function, development, and illness (Ozsolak & Milos, 2011).

Because it sheds light on the molecular mechanisms behind carcinogenesis and progression, gene expression analysis has become an essential tool in cancer research. To quantify the expression of individual genes with high sensitivity and specificity, methods like real-time quantitative PCR (RT-qPCR) are frequently employed (Bustin et al., 2009). A more comprehensive understanding of gene expression patterns is made possible by the simultaneous assessment of thousands of genes made possible by microarray technology (Schena et al., 1995; Golub et al., 1999). The genes being studied (BRCA1, PALB2, CDK1, CDH1, and CHEK2) are essential for cell cycle control, tumor suppression, and DNA repair. These genes' dysregulation leads to genomic instability, which is a characteristic of the development of cancer. For example, it is generally known that BRCA1 and BRCA2 mutations cause hereditary breast cancer, but new research shows that these mutations are expressed in sporadic instances across subtypes (Chen et al., 2022). A higher incidence of luminal breast cancer subtypes has been associated with the BRCA2-interacting protein PALB2 (Turunen et al., 2025). Cell cycle progression is driven by CDK1, which is overexpressed in proliferative subtypes such as TNBC and luminal B (Li et al., 2025). Luminal characteristics are correlated with CHEK2 mutations (Shiovitz & Korde, 2025). In invasive lobular tumors, CDH1 (which codes for E-cadherin) is often downregulated, which encourages metastasis (Kowalczyk et al., 2025).

Determining biomarkers and clarifying cancer mechanisms require an understanding of these genes' subtype-specific expression patterns. Multivariate techniques like MANOVA are required since traditional univariate analysis ignores intergene relationships. MANOVA is perfect for high-dimensional genomic data because it evaluates differences in several dependent variables between groups (Chen et al., 2009; Cheng et al., 2024).

Materials and Methods

The data-gathering process and the statistical methods used for data analysis are the main topics of this section. A one-way multivariate analysis of variance approach using model assumptions was covered in this section. The post hoc multiple comparisons were also covered.

Survey Design

The secondary dataset consisted of 1,139 breast cancer patients sourced from the University of Calabar Teaching Hospital, Calabar. The dataset includes categorical breast cancer subtypes HER2 (Human Epidermal Growth Factor Receptor 2), Luminal A, Luminal B, and Triple-Negative Breast Cancer (TNBC) as the independent variable, while the continuous gene expression levels of BRCA1 (Breast Cancer 1), PALB2 (Partner and Localizer of BRCA2), CDK1 (Cyclin-Dependent Kinase 1), CDH1 (Cadherin 1/E-cadherin), and CHEK2 (Checkpoint Kinase 2) served as the dependent variables.

Model Assumptions

To guarantee the validity of the statistical process, the basic presumptions of the general linear model were evaluated before performing the MANOVA.

Multicollinearity

When one dependent variable may be described as a linear combination of other dependent variables, multicollinearity occurs, inflating the correlations between the variables and possibly skewing the results. The Variance Inflation Factor (VIF) was used in this study to evaluate multicollinearity. The VIF gives a measure of how much collinearity increases the variance of an estimated regression coefficient. It is interpreted using standard thresholds:

If VIF = 1, there is no evidence of multicollinearity because there is no correlation between the predictors.
ii. moderate correlation that is normally acceptable is indicated by $1 \leq VIF \leq 5$

Univariate Normality Assumption:

According to the univariate normality assumption, every variable in the dataset must have a normal distribution, meaning that the majority of the observations are concentrated close to the mean and the data points are symmetrically dispersed around it. The Shapiro-Wilk test was used in the current investigation to evaluate this assumption. The Shapiro-Wilk test determines if a random sample x_i, i = 1, 2,..., n, is taken from a Gaussian distribution with true variance and mean, respectively.
The following is the formulation of the hypotheses:
Ho: The distribution of the random sample is normal.
H_1: There is no normal distribution in the random sample.

The definition of the Shapiro-Wilk test statistic is:

$$W = \frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where x_i are the ordered sample values and a_i are constants determined by:

$$(Q_1, Q_2 \ldots Q_i) = \frac{m^t v^{-1}}{(m^t v^{-1} m)^{1/2}},$$

with v standing for the order statistics' covariance matrix and M = (M_{11}, M_{22}, ----, M_{nn}) for the predicted values of the ordered statistics.

If the p-value is less than 0.05, the null hypothesis of normality is rejected, and it is determined that the sample does not have a normal distribution.

Multivariate Normality Assumption

The combined distribution of the dependent variables must be roughly normal within each level of the independent grouping variable to meet the multivariate normality assumption. The Mahalanobis distance was used in the current investigation to evaluate this assumption. By taking into consideration the relationships between variables, the Mahalanobis distance calculates how far an observation is from the mean of a multivariate distribution. The Mahalanobis distance is a reliable metric for identifying multivariate outliers and deviations from normalcy because it takes into account the dataset's covariance structure, in contrast to the Euclidean distance, which handles each variable separately.

The formal expression for the Mahalanobis distance is: $D^2 = (x - m)^t c^{-1} (x - m)$,

where; D^2 denotes the squared Mahalanobis distance,

x is the vector of observed values,

m is the vector of mean values for the variables, and

c^{-1} represents the inverse of the covariance matrix of the variables.

Test for Equality of Covariance Matrices

The covariance structure of the dependent variables must be constant across all groups in order for the assumption of equality of covariance matrices to hold. Formally, the covariance matrix is taken to be constant for each observation vector yi, where (i = 1, 2,..., n): cov(yi) = ∑, where ∑ is the common covariance matrix, represented in matrix form

$$\text{a} \quad \text{s} \quad : \quad cov(y_i) = \sum \quad = \quad \begin{pmatrix} \sigma_{11} & \sigma_{12} & - & - & - & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & - & - & - & \sigma_{2p} \\ : & : & & & & : \\ \sigma_{p1} & \sigma_{p2} & - & - & - & \sigma_{pp} \end{pmatrix}$$

Box's M test of equality of covariance matrices was used in this work to assess the assumption. This test officially investigates the homogeneity of the dependent variables' covariance matrices across groups.
The following theories are put to the test:
Ho: The dependent variables' covariance matrices are the same for all groups.

H1: The dependent variables' covariance matrices differ between groups.
The expression for the test statistic is:

$$U = \left(\sum_i \frac{1}{(ni-1)} - \sum_i \frac{1}{(ni-1)}\right)\left(\frac{2p^2+3p-1}{6(p+1)(g-1)}\right)$$

where g is the number of groups and p is the number of dependent variables. If the significance level is $p < 0.05$, the null hypothesis Ho is rejected.

Test for Equality of Error Variance

This test determines if the variance of residuals (errors), a quality known as homoscedasticity, is constant across several groups or levels of the independent variable. The validity of many statistical analyses, such as regression and ANOVA, can be impacted by identifying heteroscedasticity, or unequal variances.
This assumption was assessed in this study using Levene's test.
Theories:
Ho: Every group's variances are the same.
H1: Not every group experiences the same deviations.

Test Statistic

The test statistic for Levene's test is given by: $\frac{(n-k)}{(k-1)} \frac{\sum_{i-1}^{k} ni\,(zi-z..)^2}{\sum_{i=1}^{k}\sum_{j=1}^{ni}(zij-zi)^2}$

where

n = total sample size

k = number of groups

n_i = sample size of the i^{th} group

z_{ij} = transformed value of the jth observation in the ith group (typically absolute deviation from the group median or mean)

z_i = mean of z_{ij} in the ith group

z = overall mean of all z_{ij}

Decision Rule

If p is less than 0.05, reject Ho. Determine that there is a significant difference between at least one pair of group variances.

Interpretation

The assumption of identical error variances (homoscedasticity) is supported if Ho is not rejected. Rejecting Ho violates the premise and may have an impact on the validity of subsequent statistical studies by demonstrating unequal error variances.

Multivariate One-Way Analysis of Variance

A statistical method for comparing the means of two or more groups across several dependent variables at once is called multivariate analysis of variance, or MANOVA. Because it takes these linkages into account while evaluating group differences, MANOVA is especially helpful when the dependent variables are associated. MANOVA assumes that p-variate normal populations with a common covariance matrix, Σ, are the source of k independent random samples, each of size n. Table 1 illustrates a one-way MANOVA design.

Table 1: One-way MANOVA design

	Sample 1		**Sample 2**	**Sample k**
	From Np (μ_1, $\sum$)		**From Np (μ_2, $\sum$)**	**From Np (μ_k, $\sum$)**
	Y_{11}	……..	Y_{21} …..…	Y_{K1}
	Y_{12}	……	Y_{22} ….....	Y_{K2}
	⋮		⋮	⋮
	Y_{1n}	……	Y_{2n} . . . . . ….	Y_{Kn}
Total	$Y_{1.}$	……	Y_2 ……...	$Yk_.$
Mean	$\overline{Y}_{1.}$	……	$\overline{Y}_{2.}$ . . . . . …	$\overline{Y}_{k.}$

The total means are defined as follows:

Total of i[th] sample: $y_i = \sum_{j=1}^{n} y_{ij}$

Overall total: $y.. = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$

Mean of the i[th] sample: $\bar{y}_{i.} = \frac{y_{i.}}{ni}$

Overall mean $\bar{y}.. = \frac{y..}{kn}$

The model for each observation vector is given as;

$y_{ij} = \mu + \alpha_i + \sum_{ij} = \mu_i + \sum_{ij}$

$i = 1,2,\ldots\ldots,k \qquad j = 1,2,\ldots,n$

where

y_{ij} = the i^{th} treatment effect due to the j^{th} observation.

μ = overall mean

α_i = the i[th] treatment effect

$\sum_{ij}$ = the error term of the model.

Test for Multivariate Effect

Several test statistics can be used to evaluate the independent variable's multivariate impact on a collection of dependent variables. Pillai's Trace is a frequently used statistic that is very resilient to violations of assumptions like equality of covariance matrices and multivariate normality.
Pillai's Trace assesses whether the combined dependent variables are statistically significantly impacted by group membership. The following is the formulation of the test's hypotheses:

The test is given by:

$$V^{(s)} = tr\left[(E+H)^{-1}H\right] = \sum_{i=1}^{s} \frac{\lambda i}{1+\lambda i}$$

Hypothesis:
H_o: There is no multivariate effect of breast cancer subtypes on the total levels of gene expression.
H_1: The total levels of gene expression are influenced by the multivariate effect of breast cancer subtypes.
A multivariate effect is demonstrated by a substantial Pillai's Trace value, which shows that at least one of the dependent variables is impacted by group differences.

Test of Between-Subject Effects

Examining the impact of the independent variable on each dependent variable separately is crucial after the multivariate test of equality of mean vectors outlined in Section **2.3.1**. The group means for each of the p dependent variables are compared using the between-subjects effects test.

Formally, equality of means for every dependent variable is implied by the overall MANOVA hypothesis of equality of mean vectors:

H_o m$_1$ = m$_2$ = L = m$_k$, This implies equality of means for each of the p variables

$$H_{0r}: \mu_{1r} = \mu_{2r} = \mu_{3r} = \cdots = \mu_{kn}, \ for \ r = 1,2, \cdots, p$$

An ordinary univariate ANOVA F-test might be used to test this hypothesis for each variable separately. The method enables researchers to pinpoint the precise dependent factors that contribute to the multivariate effect that the overall MANOVA test detects.

Post Hoc Test of Multiple Comparisons

In order to control the experiment-wise Type I error rate, post-hoc tests are used to investigate variations between different group means. Once a significant ANOVA result has been obtained, they are utilized to determine which particular group means differ considerably. Scheffé's test was used in this investigation because it is appropriate for comparisons with different group sizes.

After an ANOVA, Scheffé's test is a conservative method that identifies the pairings of means where there are significant differences. The test compares every feasible set of group means. If there are three group means, for instance, the pairwise comparisons consist of:

After the ANOVA procedure is completed, the Scheffe test is used to identify the location of the significant difference in the means. The Scheffe test requires comparing the means two at a time using every feasible mean combination. For instance, the following comparisons must be made if there are three means: $\underline{X}_1$ versus $\underline{X}_2$, $\underline{X}_1$ versus $\underline{X}_3$, $\underline{X}_2$ versus $\underline{X}_3$.

The Scheffe's test statistic is given as:

$$F_s = \frac{(\underline{xi} - \underline{x}_j)^2}{\delta_w^2\left(\frac{1}{ni}\right) + \left(\frac{1}{nj}\right)},$$

where;

$\underline{x}_i$ and $\underline{x}_j$ are the means of the samples being compared,

n_i and n_j are the respective sample sizes,

σ_w^2 is the within-group variance. A significant F_s value indicates a statistically significant difference between the corresponding group means, allowing for identification of which groups contribute to the overall effect observed in the ANOVA.

RESULTS AND DISCUSSION

Results

Table 2: Test For Multicollinearity.

Dependent variable	VIF
BRCA1	1.648
PALB2	1.012
CDK1	1.063
CDH1	1.491
CHEK2	1.172

Table 3: Shapiro-Wilk Test of Normality

Dependent variable	Test value	p-value
BRCA1	0.63	.043
PALB2	0.53	.023
CDK1	0.47	.017
CDH1	0.50	.020
CHEK2	0.23	.043

Table 4: Mahalanobis Test

Test statistic	*Minimum*	*Maximum*
Mahalanobis Distance	.254	12.679

Table 5: Box's M Test

(*Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups*)

Test Value	*F-Ratio*	*Df1*	*Df2*	*P-Value*	*Decision (α=0.05)*
131.883	2.897	45	658295.469	.000	Reject H_o

Table 6: Levene's Test of Equality of Error Variances

(*Tests the null hypothesis that the error variance of the dependent variable is equal across groups*).

Dependent variable	Levene Statistic	p-value	Decision (α =0.05)
BRCA1	2.424	.089	Reject H_o
PALB2	3.341	.036	Accept H_o
CDK1	2.455	.086	Reject H_o
CDH1	.914	.401	Reject H_o
CHEK2	1.013	.363	Reject H_o

Table 7: Between-Subjects Factors

Cancer Subtypes	Count
HER2	161
LUMINAL A	548
LUMINAL B	315
TNBC	115
Total	1139

Table 8: Multivariate Test

Test statistic	Test Value	F-Ratio	p-value	Observed Power	Decision (α = 0.05)
Pillai's Trace	.525	48.085	.000	1.000	Reject H_o
Wilks' Lambda	.522	55.304	.000	1.000	Reject H_o
Hotelling's Trace	.828	62.352	.000	1.000	Reject H_o
Roy's Largest Root	.709	160.735	.000	1.000	Reject H_o

Table 9: Tests of Between-Subjects Effects

Dependent Variable	Df	Mean Square	F-value	p-value.	Observed Power
BRCA1	3	30.648	54.250	.000	1.000
PALB2	3	4.300	4.590	.003	.890
CDK1	3	113.798	258.419	.000	1.000
CDH1	3	15.638	28.255	.000	1.000
CHEK2	3	48.796	84.770	.000	1.000

Table 10: Scheffe Test of Multiple Comparisons

Dependent Variable	Cancer Subtypes	Cancer Subtypes	P-Value	*Decision* ($\alpha = 0.05$)
BRCA1*	HER2*	LUMINAL A	.000	Has effect
		LUMINAL B	.000	Has effect
		TNBC	.887	No effect
	LUMINAL A*	HER2	.000	Has effect
		LUMINAL B	.000	Has effect
		TNBC	.018	Has effect
	LUMINAL B*	HER2	.000	Has effect
		LUMINAL A	.000	Has effect
		TNBC	.000	Has effect
	TNBC*	HER2	.887	No effect
		LUMINAL A	.018	Has effect
		LUMINAL B	.000	Has effect
PALB2*	HER2*	LUMINAL A	.436	No effect
		LUMINAL B	.010	Has effect
		TNBC	.139	No effect
	LUMINAL A*	HER2	.436	No effect
		LUMINAL B	.094	No effect
		TNBC	.611	No effect
	LUMINAL B*	HER2	.010	Has effect
		LUMINAL A	.094	No effect
		TNBC	.987	No effect
	TNBC*	HER2	.139	No effect
		LUMINAL A	.611	No effect
		LUMINAL B	.987	No effect
CDK1*	HER2*	LUMINAL A	.000	Has effect
		LUMINAL B	.513	No effect
		TNBC	.000	Has effect
	LUMINAL A*	HER2	.000	Has effect
		LUMINAL A	.000	Has effect
		TNBC	.000	Has effect
	LUMINAL B*	HER2	.513	No effect
		LUMINAL A	.000	Has effect
		TNBC	.009	Has effect

	TNBC*	HER2	.000	Has effect
		LUMINAL A	.000	Has effect
		LUMINAL B	.009	Has effect
Dependent Variable	**Cancer Subtypes**	**Cancer Subtypes**	**P-Value**	***Decision* (α = 0.05)**
CDH1*	HER2*	LUMINAL A	.000	Has effect
		LUMINAL B	.410	No effect
		TNBC	.000	Has effect
	LUMINAL A*	HER2	.000	Has effect
		LUMINAL B	.000	Has effect
		TNBC	.979	No effect
	LUMINAL B*	HER2	.410	No effect
		LUMINAL A	.000	Has effect
		TNBC	.000	Has effect
	TNBC*	HER2	.000	Has effect
		LUMINAL A	.979	No effect
		LUMINAL B	.000	Has effect
CHEK2*	HER2*	LUMINAL A	.000	Has effect
		LUMINAL B	.603	No effect
		TNBC	.000	Has effect
	LUMINAL A*	HER2	.000	Has effect
		LUMINAL B	.000	Has effect
		TNBC	.000	Has effect
	LUMINAL B*	HER2	.603	Has effect
		LUMINAL A	.000	Has effect
		TNBC	.000	Has effect
	TNBC*	HER2	.000	Has effect
		LUMINAL A	.000	Has effect
		LUMINAL B	.000	Has effect

DISCUSSION

The findings of this study provide a comprehensive and lucid picture of how gene expression patterns vary among molecular subtypes of breast cancer. This offers fresh perspectives and validates many of the theories proposed in the literature on molecular oncology. The dataset's statistical soundness was confirmed by the initial diagnostic tests. For every gene under study, the Variance Inflation Factors (VIFs) (**Table 2**) were less than 5, suggesting that multicollinearity was not a problem. This is in line with Stevens' (2009) observation that MANOVA results may be skewed if variables have excessive correlations. The majority of genes showed approaching univariate normalcy according to the Shapiro-Wilk test (**Table 3**), except BRCA1 and CHEK2. This bolsters the claim made by Law et al. (2014) that because of their biological variability, gene expression data frequently defy strict normalcy. The absence of multivariate outliers in Mahalanobis distances (**Table 4**) enhances the analysis's dependability. However, Box's M test (**Table 5**) revealed a breach of covariance matrix

homogeneity, which can affect sensitivity, as Olson (1976) cautioned. However, this work adhered to methodological best practices suggested by Warne (2014) by using Pillai's Trace, which is known to be robust to such violations.

The MANOVA results, which form the basis of the findings, showed that the expression of the five genes under investigation BRCA1, PALB2, CDK1, CDH1, and CHEK2 is significantly influenced by the subtypes of breast cancer. With observed power at 100%, the multivariate test statistics Pillai's Trace, Wilks' Lambda, Hotelling's Trace, and Roy's Largest Root (**Table 8**) were all very significant. This supports the main assertion made by Perou et al. (2000) that molecular subtypes are biologically unique entities with coordinated variations in gene expression. These findings further support the claim made by Langfelder and Horvath (2008) that multivariate models are superior to univariate approaches in capturing interdependencies in genomic data. The significance of specific genes was further highlighted by the tests of between-subject effects (**Table 9**). Significant variation was seen in BRCA1, CDK1, CDH1, and CHEK2 among subtypes; this seems to be in line with what is now known about their roles in the development and spread of tumors. For example, Hanahan and Weinberg's (2011) discovery that DNA repair gene disruptions are essential to carcinogenic pathways is supported by the increased impact of BRCA1 in luminal and HER2 subtypes.

Its description as a proliferation driver in aggressive phenotypes is supported by CDK1, which is significantly important in HER2 and TNBC (Li et al., 2025). Recent results that relate CHEK2 mutations to luminal phenotypes (Shiovitz & Korde, 2025) are consistent with the importance of CHEK2 in luminal A, HER2, and TNBC. Contrary to previous links with luminal malignancies, PALB2 did not exhibit significant subtype variation (Turunen et al., 2025). This implies that compared to familial risk studies, its impact might be less pronounced in population-based datasets.

The subtle variations among subtypes were revealed by the post-hoc Scheffe tests (Table 10). As suggested by the basal-like classification of TNBC (Chen et al., 2025), triple-negative cancers may rely more on other genomic instabilities and less on BRCA1 dysregulation. BRCA1 expression significantly diverged between luminal A, luminal B, and HER2, but not TNBC. Luminal A, HER2, and TNBC were reliably separated by CDK1, CDH1, and CHEK2, highlighting the significance of proliferation, adhesion loss, and checkpoint dysregulation in characterizing aggressive behavior in these subtypes. Luminal B exhibited the weakest gene-level differentiation in this dataset, despite being clinically recognized as an intermediate-risk subtype (Guo et al., 2025). This suggests that its biology may be more diverse or defined by proliferative markers like Ki-67 rather than the five genes examined here.

These findings collectively demonstrate that, of the chosen indicators, luminal A and HER2 subtypes had the greatest impact on gene expression, closely followed by TNBC. The results, which highlight the variety even within categories like TNBC, are highly consistent with recent genomic revisions of subtype classification (Laenkholm et al., 2025). Furthermore, the strong multivariate effects found here empirically support the clinical implications of intrinsic molecular subtypes, where treatment choices are based on the underlying biology: experimental immunotherapy or DNA repair targeting strategies for TNBC, anti-HER2 agents for HER2-enriched tumors, and hormone therapy for luminal types.

The MANOVA results confirm the importance of molecular subtyping in breast cancer as a reflection of coordinated gene expression dynamics and as a therapeutic aid. This study advances our knowledge of the genetic architecture underlying breast cancer heterogeneity by revealing subtype-specific effects in BRCA1, CDK1, CDH1, and CHEK2, while indicating

modest relevance of PALB2. The findings support the findings of Sung et al. (2021) about the worldwide impact of breast cancer, but they also go one step further by linking those epidemiological findings to the molecular interactions that influence prognosis and treatment response.

CONCLUSION

The analysis's results highlight the molecular heterogeneity of breast cancer by showing a distinct pattern of gene expression variations specific to molecular subtypes. Although the normality and homogeneity assumptions were not entirely satisfied, the preliminary diagnostic tests verified that the data were mainly free of multicollinearity and multivariate outliers, necessitating cautious interpretation. Notwithstanding these drawbacks, the multivariate tests yielded strong evidence that the collective expression of BRCA1, PALB2, CDK1, CDH1, and CHEK2 is significantly influenced by breast cancer subtypes; observed power reached 100%, confirming the model's suitability.

Additional univariate testing of inter-subject effects showed that PALB2 exhibited the weakest, yet statistically significant, differentiation among the breast cancer subtypes ($p = 0.003$), whereas BRCA1, CDK1, CDH1, and CHEK2 demonstrated bigger subtype-specific differences. Luminal A and HER2 subtypes consistently showed strong effects across most genes, but TNBC showed selective influences and luminal B remained somewhat less distinctive, according to post hoc comparisons. Together, these findings show that luminal B exhibits weaker subtype-defining gene expression patterns, TNBC reflects its well-known variability, and luminal A and HER2 subtypes maintain more prominent and consistent gene expression profiles.

In the end, the analysis confirms the significance of multivariate methods in elucidating the complex genetic foundations of breast cancer, highlighting the roles of specific genes in differentiating subtypes and providing insights that could guide future precision-based treatment approaches.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Consent for publication: Not applicable.

Data Availability Statement: The data used in the current study are available from the corresponding author and can be attached on request

Conflicts of Interest: The authors declare no conflicts of interest

REFERENCES

Alghamdi, N. A., Alharbi, A. S., Alqahtani, A. S., & Alghamdi, S. A. (2025). Prognostic significance and molecular classification of triple-negative breast cancer: A systematic review. *European Journal of Breast Health, 21*(2),101–114. https://doi.org/10.4274/ejbh.galenos.2024.2024-10-2

Alqahtani, A. J., Alkayyal, M. A., Alshammari, F. D., Alshammari, A. A., Alshammari, A. A., & Alshammari, A. A. (2025). Molecular subtypes and mechanisms of breast cancer: Precision medicine approaches for targeted therapies. *Cancers,17*(7),1102. https://doi.org/10.3390/cancers17071102

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCRexperiments.*ClinicalChemistry,55*(4),611–622. https://doi.org/10.1373/clinchem.2008.112797

Chen, J., Wang, Y., & Li, L. (2025). Intrinsic subtype and immunity score in identification of triple-negative breast cancer at low risk. *The Breast, 80*, 102–110. https://doi.org/10.1016/j.breast.2025.01.018

Chen, L., Chiquet, J., Bonneau, R., & Kirk, P. (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics, 25*(7), 897–903. https://doi.org/10.1093/bioinformatics/btp098

Chen, Y.-C., Lin, Y.-F., Chang, W.-C., Lin, C.-F., & Wang, P.-H. (2022). High BRCA1 gene expression increases the risk of early distant metastasis in ER+ breast cancers. *Scientific Reports, 12*(1), Article 77. https://doi.org/10.1038/s41598-021-03471-w

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Guo, Y., Zhang, Y., & Li, X. (2025). Drug resistance in breast cancer: Mechanisms and strategies for management. Drug Resistance Updates, 83, Article101288. https://doi.org/10.1016/j.drup.2025.101288

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hotelling, H. (1931). The generalization of Student's ratio for use with two or more than two variates. *Biometrika, 23*(3/4), 463–485. https://doi.org/10.1093/biomet/23.3-4.463

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). Wiley.

Huberty, C. J., & Petoskey, M. D. (2000). Multivariate analysis of variance and covariance. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 183–227). Academic Press. https://doi.org/10.1016/B978-012688290-5/50007-3

Kowalczyk, K., Kowalczyk, A., & Kowalczyk, M. (2025). The role of the CDH1 gene in the pathogenesis and progression of lobular breast cancer. *Cureus, 17*(8), Article e89290. https://doi.org/10.7759/cureus.89290

Laenkholm, A.-V., Grabau, D., & Kvist, P. (2025). The Breast Cancer Classifier refines molecular breast cancer classification to delineate the HER2-low subtype. *npj Breast Cancer, 11*(1), Article 19. https://doi.org/10.1038/s41523-025-00723-0

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics, 9*, Article 559. https://doi.org/10.1186/1471-2105-9-559

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*(2), P14. https://doi.org/10.1186/gb-2014-15-2-r14

Li, X., Wang, Y., & Zhang, L. (2025). SENP5 drives breast cancer progression through deSUMOylatingofcdk1.*BreastCancerResearch,27*(1),Article100. https://doi.org/10.1186/s13058-025-02054-1

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin, 83*(4), 579–586. https://doi.org/10.1037/0033-2909.83.4.579

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics, 12*(2), 87–98. https://doi.org/10.1038/nrg2901

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Chen, Y., & Jeffrey, S. S. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747–752. https://doi.org/10.1038/35021093

Rossing, M., Tvedskov, T. F., & Ejlertsen, B. (2021). Clinical implications of intrinsic molecular subtypes of breast cancer for sentinel node status. *Scientific Reports, 11*(1), Article 2259. https://doi.org/10.1038/s41598-021-81750-4

Santonja, A., Mancikova, V., & Castaño, J. P. (2024). Clinical characteristics and prevalence of molecular subtypes in breast cancer: A single-center experience. *Journal of Clinical Oncology, 42*(suppl 16), e13045. https://doi.org/10.1200/JCO.2024.42.16_suppl.e13045 (Note: Adapted from related search results for closest match.)

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Shiovitz, S., & Korde, L. A. (2025). CHEK2-related cancer predisposition. In M. P. Adam, H. H. Ardinger, B. Feldman, et al. (Eds.), *Gene Reviews®*. University of Washington. Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249. https://doi.org/10.3322/caac.21708

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Turunen, H., Kaikkonen, M. U., & Ylä-Herttuala, S. (2025). PALB2 mutations increase oncogenic properties of breast epithelial cells by enhancing JAM3 and PARVB expression. Biochemical and Biophysical Research Communications, 728, Article 150345. https://doi.org/10.1016/j.bbrc.2025.150345

Wang, Y., Li, X., & Zhang, Y. (2025). Classifications of triple-negative breast cancer: Insights and current therapeutic approaches. *Cell & Bioscience, 15*(1), Article 13. https://doi.org/10.1186/s13578-025-01359-0

Warne, R. T. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research & Evaluation, 19*(17), 1–10.

Wilks, S. S. (1932). Certain remarks concerning the comparison of several experimental population means. *Biometrika, 24*(3/4), 419–426. https://doi.org/10.1093/biomet/24.3-4.419

Yershova, K., & Singhal, S. (2022). Subtypes of breast cancer. In *Stat Pearls*. Stat Pearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK583808/

Yin, L., Duan, R., Bian, X.-W., & Yu, L. H. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research, 22*(1), Article 61. https://doi.org/10.1186/s13058-020-01296-5/.