Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 3
BLSME 2024

# Bioinformatic Insights into the Challenges of miRNA-Based BRCA Status Classification

Haoyue Jia *

School of Life Sciences, Sun Yat-sen University, Guangzhou, China

* Corresponding Author Email: jiahy6@mail2.sysu.edu.cn

**Abstract.** BRCA1 and BRCA2 are vital tumor-suppressed genes closely related to the presence of ovarian cancer. Recent bioinformatic approach has been applied for identifying the potential carrier of BRCA mutant via miRNA microarray expression. However, our statistical analysis suggests that the limma filter method may not reflect true genotype status despite the significance some data render. In order to refine the identification of the BRCA subtype through differentially expressed miRNAs, the rank sum ratio method was employed to identify miRNAs demonstrating significance in two distinct tasks. The outcome highlights miRNAs that are not included in previous limma filtration for BRCA identification, suggesting an innovative perspective when analysing the expression pattern for disease genotype recognition.

**Keywords:** Limma analysis; differential expressed miRNA; microarray; BRCA.

## 1. Introduction

Cancer is a time-dependent disease with DNA degeneration and impairness constantly accumulated in the human body. Normally, proto-oncogenes are inactivated, and tumor-suppressed genes are activated during non-cancerous genetic inheritance, while corresponding harmful variants could be carried and inherited through distinctive racial and geographic populations, rendering them more prone to certain cancer diseases. BRCA1 and BRCA2, whose defection are embryonically lethal, are typical tumor-suppressed genes functioning in homologous recombination of DNA repair system [1]. Inherited defects of BRCA1 and BRCA2 strongly relate to the cancer predisposition in breast and ovarian and also do harm to women brain integrity through reduction of gray matter [2]. For those diagnosed as BRCA mutants, prophylactic bilateralsalpingo-oophorectomy, namely the removal of ovaries is usually recommended prior to the formation of ovarian cancer, which, though mostly effective, could also increase the risk of late life dementia and Alzheimer's disease [3-4]. Identification of BRCA1 and BRCA2 mutation is thus vital in preventing cancer from making presence for these carriers to reduce the risk.

Health care and genetic counsel are usually available for anyone who is concerned to have harmful BRCA1 and BRCA2 gene variants through routine genetic diagnose by their family history. However, such dogmatic test approach has also been gradually undesirable for United States Preventative Service Task Force since only 0.2-0.3% of general women population in average are estimated to carry BRCA mutations during this process [5]. Hence, bioinformatic approach is also getting spring up trying to evaluate the BRCA mutants by genome-wide microarray expression analysis recently, which is a high-throughput technology used for detecting gene expression levels by performing gene wise local hybridization with distinctive RNA counts measurements. BRCA mutation identification by miRNA microarray design is showing up in recent bioinformatic research, but scarce has gone further to separate BRCA1 and BRCA2 mutation by their miRNA expression pattern [6].

Besides, various types of methods for finding differential expressed genes in microarray experiment are developed for their convenience and accuracy, among which limma, or Linear Models for MicroArray and RNA-seq Data available from the R packages has taken a rather popular place [7]. However, whether the abuse use of limma, which is supposed to guide the downstream gene

enrichment process by traditional bioinformatic research, could successfully apply in deduction and clustering of the genotype identification in clinical practice remains a question.

In this study, the detailed separation of BRCA1 and BRCA2 deficiency would be attempted and identification of BRCA genotype would be applied by cluster analysis. It could be then demonstrated that the differential expressed miRNAs filtered by limma is not necessarily reliable on genotype identification and possible reasons would be introduced.

## 2. Data and Methods

### 2.1. Dataset

Microarray data of the genome-wide miRNA in wildtype and BRCA mutated samples had been obtained from the Gene Expression Omnibus database of the National Centers for Biotechnology Information as GSE226445. The GEO series contain 227 miRNAs and 653 samples from six independent international cohort, with 132 samples from University of PenniSylvia are used for calibration. The other 521 samples are used to find differential expressed miRNAs. The study population characteristics are summarized in Table 1.

**Table 1.** Statistical characteristics of the studied cohort samples

<table>
<tr><td>Cohort</td><td colspan="3">Samples (N = 653)</td></tr>
<tr><td></td><td>Wild-type</td><td>BRCA mutated</td><td>No ovaries at test</td></tr>
<tr><td>UPenn</td><td>83</td><td>49</td><td>0</td></tr>
<tr><td>DFCI</td><td>200</td><td rowspan="2">0</td><td rowspan="2">0</td></tr>
<tr><td>DGO</td><td>20</td></tr>
<tr><td>CCGP</td><td>0</td><td>162</td><td rowspan="3">75</td></tr>
<tr><td>BWH</td><td>0</td><td>87</td></tr>
<tr><td>IHCC</td><td>0</td><td>52</td></tr>
<tr><td rowspan="3">Total</td><td rowspan="3">83+220</td><td>Unidentified</td><td>49 + 139</td></tr>
<tr><td>BRCA 1</td><td>79</td></tr>
<tr><td>BRCA 2</td><td>83</td></tr>
</table>

### 2.2. Methods

#### 2.2.1. Limma regression analysis

Limma linear model is often constructed in microarray analysis when there are two or more explanatory variables that are of interest. Mean model and mean reference model are then applied in different scenarios to comply with downstream analysis, each corresponding to distinctive design matrix, without or with intercept [8]. In mean model, factorial levels of certain factors are built with equal numbers of parameters to estimate. In mean reference model, certain combination of different factor levels is designed as the reference level. The rest of the levels are parameterized relative to the reference. Once the difference of certain level is added towards the mean reference, corresponding measurement moves towards such marginal level with other factors unchanged geometrically. In the model with regard of this research, the limma model is built with ~0 + BRCAness + Ovarystatus, or:

$$y = \beta_1 x_{WT} + \beta_2 x_{MT} + \beta_3 x_{no\_ovary} \quad (1)$$

The construction of BRCA genotype is framed by mean model, giving two parameters to estimate respectively, and another parameter indicating the presence of prior bilateral salpingo-oophorectomy operation treatment (denoted as T hereinunder), which is additive to build a mean reference model on the whole. Corresponding parameter is denoted by $\beta_1$, $\beta_2$ and $\beta_3$ as regression coefficients. Similarly, binary indicator variables each value 1 given the specification of wildtype, mutated type

and occurrence of treatment T and 0 if not is denoted as $x_{i1}$, $x_{i2}$ and $x_{i3}$, with their row index referring to the sample order. Evidently, the occurrence of T is a sufficient condition for the deduction of mutated sample but not vice versa. In this simple case, the expression of each estimator could be expressed following with demonstration hereinunder.

For the regression model, we denote $D$, $D^T$ as design matrix and its transpose, which is essential for storing the value of explanatory variables [9]. $E$ the column vector of expression as the response variable $y_i$. The denotation of each element in $D^T D$ could also be applied for further clear algebraic manifestation.

$$\begin{bmatrix} \sum_i & x_{i1}^2 \sum_i & x_{i1}x_{i2} \sum_i & x_{i1}x_{i3} \sum_i & x_{i1}x_{i2} \sum_i & x_{i2}^2 \sum_i & x_{i2}x_{i3} \sum_i & x_{i1}x_{i3} \sum_i & x_{i2}x_{i3} \sum_i & x_{i3}^2 \end{bmatrix} = \begin{bmatrix} X & A & B & A & Y & C & B & C & Z \end{bmatrix} \tag{2}$$

Hence, the estimation of regression coefficients is calculated by:

$$\begin{aligned} \hat{\beta} &= (D^T D)^{-1} D^T E \\ &= \begin{vmatrix} X & A & B \\ A & Y & C \\ B & C & Z \end{vmatrix}^{-1} \begin{bmatrix} \sum_i (YZ - C^2) x_{i1} y_i + (BC - AZ) x_{i2} y_i + (AC - BY) x_{i3} y_i \\ \sum_i (BC - AZ) x_{i1} y_i + (XZ - B^2) x_{i2} y_i + (AB - XC) x_{i3} y_i \\ \sum_i (AC - BY) x_{i1} y_i + (AB - XC) x_{i2} y_i + (XY - A^2) x_{i3} y_i \end{bmatrix} \end{aligned} \tag{3}$$

For the condition given aforementioned, it is then easily deducted that $A$ and $B$ both equals zero and $C$ equals the sample size with treatment T, since genotype is mutually exclusive and T only occurs within the subset of mutated type. The value of $X$, $Y$ and $Z$ are always identical to the sample size of wildtype, mutated type and T regardless of the distribution of T within the cohort, each and their corresponding set denoted as $WT$, $MT$ and $N$. Therefore, the determinant of $D^T D$ would be $NMTWT - WTN^2$. The expression of wildtype and mutated is then simplified as

$$\hat{\beta}_1 = \frac{1}{det(D^T D)} \left( \sum_{i \in WT} (MTN - N^2) y_i \right) = \frac{1}{WT} \sum_{i \in WT} y_i \tag{4}$$

$$\hat{\beta}_2 = \frac{1}{det(D^T D)} \left( - \sum_{j \in MT} (WTN) y_j + \sum_{k \in N} (WTN) y_k \right) = \frac{1}{MT - N} \sum_{j \in MT - N} y_j \tag{5}$$

With The difference set $MT - N$ is defined as mutated sample set without T.

Noticeably, it is then worth pointing out the above result if deployed with the converse of state T in the whole cohort would stay the same. The reversed T set $N'$ together with $N$ compose a complete family of sample and is comprised of all the sample points in both wildtype set and partial mutated set. The different relationship among three sets result in a distinctive design matrix and $D^T D$, but with $D^T D$ the same determinant. In this situation, it could be obtained that:

$$\begin{aligned} \hat{\beta}_1 - \hat{\beta}_2 &= \frac{1}{det(D^T D)} \left( \sum_{i \in WT} (MTN + WTN - N^2) y_i - \sum_{j \in N'} (WTN) y_j \right) \\ &= \frac{MT + WT - N}{WT(MT - N)} \sum_{i \in WT} y_i - \frac{1}{MT - N} \sum_{j \in N'} y_j \end{aligned} \tag{6}$$

By comparison, we show that the two algebraic equations are identical in that the set $N'$ contains exactly more of a $WT$ sample points than that of the differential set $MT - N$, which generates a $(MT - N)^{-1}$ coefficient term that could precisely be absorbed by the first term of original right-hand expression of contrast $\hat{\beta}_1 - \hat{\beta}_2$. This contrast result, though fluctuates if given different 0-1 distribution of T that does not necessarily take the opposite of the original treatment design, will be

universal without the premise of T treated sample points residing in the mutated set if given delicate demonstration.

Such result shows that the contrast caused by main factors will take the heterogeneity information instead of the quantity information of the third external factor into account. Such contrast in limma only interprets the weighted average of every property-matched set. This may cause problems when it comes into categorizing real data with high variance into genotype.

Besides, since the wildtype data has no intersection with treatment T in the original dataset, it is impossible to have a complete factorial design to estimate four level parameters in limma (genotype × Ovarystatus) as well as the interaction model measuring the impact of salpingo-oophorectomy operation on certain mutation type. Building such model is prone to generate design matrices whose ranks are not full. However, it could be done when design matrix considers the subtype of BRCAness and salpingo-oophorectomy operation, with a pure mean reference model constructed as:

$$y = \beta_0 + \beta_4 x_{2-1} + \beta_5 x_{no\_ovary} + \beta_6 x_{2-1} x_{no\_ovary} \tag{7}$$

Which takes the BRCA1 patients with presence of ovary as the reference level. The model is able to estimate the effect of the interaction between the removal of ovaries and the difference between BRCA1 and BRCA2. The candidate miRNAs significantly differ in such model will be selected, with those possessing significance differing wildtype and BRCA mutation to test the identification of different genotype altogether.

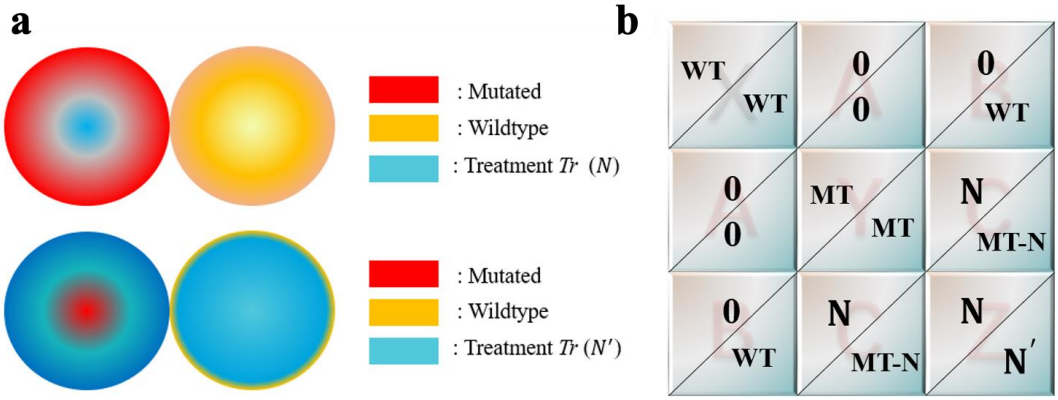


**Figure 1.** Sample genotype property related to the limma analysis

**a.** Overview of the relationship among different sample set, with the upper one indicates the original set relationship. The Treatment, or the operation of salpingo-oophorectomy is the subset of mutant group. The lower one is the putative set description where the treatment factor is reversed within the sample cohort. **b.** The matrix expression of $D^T D$ for situations in Fig.1a respectively, where WT indicates the number of wild-type samples, MT the mutated, N the ovaries removed, and N' the complement of N.

### 2.2.2. Rank sum ratio arrangement

In this study, the delicate separation of wild-type, BRCA1 and BRCA2 genotype in the sample cohort could be broken down into two independent tasks: finding differential expressed miRNAs between wild-type and BRCA, which is mainly completed by GSE226445, obtaining 19 miRNAs using the limma model ~0 + BRCAstatus + Ovarystatus, and finding differential expressed miRNAs between BRCA1 and BRCA2, each denoted as Task1 and Task2. To effectively pick out the miRNA that could do both of the work, the p values given by each limma is recorded as the support of their final score for reference. By mapping their p value linearly into the rank for all candidates and summing up all their score in each task, it has been revealed that miRNAs which could both behave well in differing wild-type from BRCA mutants and BRCA1 from BRCA2. For n candidates in total and

global maximum and minimum p value in two tasks denoted as $p_{max}$ and $p_{min}$, the linear rank mapping formula is:

$$R(p) = (n-1) \cdot \frac{p_{max}-p}{p_{max}-p_{min}} + 1 \tag{8}$$

After mapping and calculating the sum score for each miRNA candidate, the rank sum ratio would be regressed with their n-fractile and given appropriate classification of levels, detailed in model development described in the later context.

## 3. Results

### 3.1. Differential expressed miRNA selection

The original GSE226445 dataset used Combat to adjust the batch effect for calibration. During this process of Task1, UPenn samples are used for adjustment reference as well as the validation set for ROC curve. The filtration is referred to undergo three main processes: selecting miRNAs that both significantly differs from wild-type to the BRCA mutants (P<0.001) and the absolute value of log2 fold change is larger than 0.5; the log2 fold change value expressing on the same direction between original and batch-adjusted data; the ratio of fold change between raw and batch-adjusted data in the range 0.8-1.25. By performing ~0 + BRCAstatus + Ovarystatus limma analysis, the outcome number of filtered miRNAs is sequentially 57, 42, 15. The final 15 candidates are all that of the 19 miRNAs described in GSE226445. Besides, a mean model without consideration of salpingo-oophorectomy operation is also constructed and followed by the three filtration steps mentioned above as a comparison. The miRNA number outcome is correspondingly 53, 40 and 19. The intersection of these 19 miRNAs finally selected and 19 miRNAs in GSE226445 is also 15, with 12 identical with that filtered by the first limma model. The hypergeometrical test also shows high significance (P<0.001).

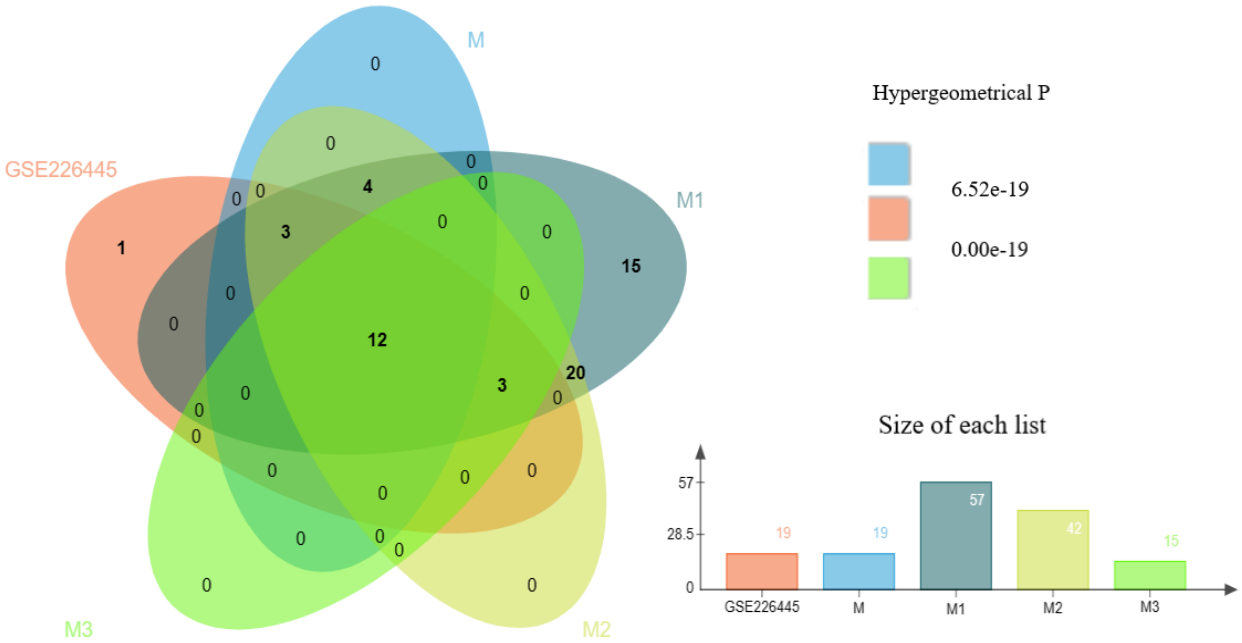


**Figure 2.** Venn diagram of five miRNAs set filtered by different model

GSE226445 for the original 19 miRNAs selected by the GEO project, M the mean model, M1, M2, M3 the ~0 + BRCAstatus + Ovarystatus limma model each after the significance filtration, same expression direction examination between original and batch-adjusted data, and the 0.8-1.25 ratio filtration respectively. Hypergeometrical test was performed between the GSE226445 and M, M3.

For Task 2, mean reference model with and without interaction term were all deployed for both raw and batch-adjusted data with clear declaration of BRCA subtype. Results show that the effects of salpingo-oophorectomy operation and its interaction on BRCA subtype are all mostly negative. Still, some miRNAs show slight reliable significance ($P<0.1$) in difference of BRCA1 and BRCA2 with interaction term within batch-adjusted data. Top 3 significances of each model are listed in Table 2. Notably, hsa-miR-18a-5p is the unique miRNA among 227 miRNAs that shows strong difference after removing ovaries ($P<0.05$), which could be given more research on relationship between BRCAness and ovary removal operation.

**Table 2.** Significance visualization with descended arrangement in each model

| Model | Statistics | p1 | p2 | p3 |
|---|---|---|---|---|
| Unbatched, no interaction | brca1-brca2 | 0.4408 | 0.4408 | 0.4408 |
| | ovaryno | 0.9184 | 0.9184 | 0.9184 |
| Batched, no interaction | brca1-brca2 | 0.0826 | 0.1499 | 0.1499 |
| | ovaryno | 0.2207 | 0.3088 | 0.3088 |
| Unbatched, interaction | brca1-brca2 | 0.8663 | 0.8663 | 0.8663 |
| | ovaryno | 0.9816 | 0.9816 | 0.9816 |
| | interaction | 0.9976 | 0.9976 | 0.9976 |
| | brca1-brca2 (ovaryno) | 0.9993 | 0.9993 | 0.9993 |
| Batched, interaction | brca1-brca2 | 0.0711 | 0.0711 | 0.0783 |
| | ovaryno | 0.0394 | 0.2363 | 0.2363 |
| | interaction | 0.9463 | 0.9463 | 0.9463 |
| | brca1-brca2 (ovaryno) | 0.3285 | 0.3285 | 0.3285 |

According to the limma analysis result and to align the candidate number, top 15 miRNAs showing strong difference between BRCA1 and BRCA2 in the batch-adjusted data with interaction term are expected to further conduct the rank sum ratio analysis. Interestingly, the 15 miRNAs selected in Task 2 has no overlap with the top 15 miRNAs found in the ~0 + BRCAstatus + Ovarystatus model, which suggest that some miRNAs are not good at distinguishing wild-type and BRCA mutants, but good at discerning BRCA1 and BRCA2 once BRCA mutant is confirmed, and for those miRNAs good at distinguishing wild-type from BRCA mutants but bad at BRCA1 and BRCA2 identification vice versa.

### 3.2. Clutser analysis

The 30 miRNA candidates merged by limma model of Task 1 and Task 2 were pooled together to give the rank sum. The rank sum ratio algorithm divides these 30 miRNAs into 3 categories: High-score candidates for 2 miRNAs: has-miR-140-5p and has-miR-142-3p, scoring 96.47 and 96.12 out of 100 respectively; low-score candidates for 8 miRNAs: hsa-miR-30d-5p, hsa-miR-19b-3p, hsa-miR-500a-3p, hsa-miR-126-5p, hsa-miR-485-3p, hsa-miR-20b-5p, hsa-miR-4433a-3p, hsa-miR-4433b-5p, with the highest score of 64.87 and lowest 54.44 out of 100; medium-score candidate for the remaining 20 miRNAs.

After the classification, three groups were applied with K-means clustering for $K = 3$, with reference of real genotype of 220 wildtype and 162 BRCA subtype confirmed mutants in the cohort. The originally selected 19 miRNAs in GSE226445 and 15 most significant miRNAs, which are mutually exclusive, were also used for clustering as comparison. The Hamming distance between the genuine genotype and high scoring group, medium scoring group, low scoring group, differential expressed (DE) miRNAs in GSE226445 and Task 2-significant group was 223, 161, 171, 157, 162 respectively, under case where the sample size is 382 (220 + 79 + 83). Unsupervised hierarchical cluster is performed on the left side of the heatmap, showing that high scoring group and Task 2-significant groups are closer to the real BRCA genotype distribution.

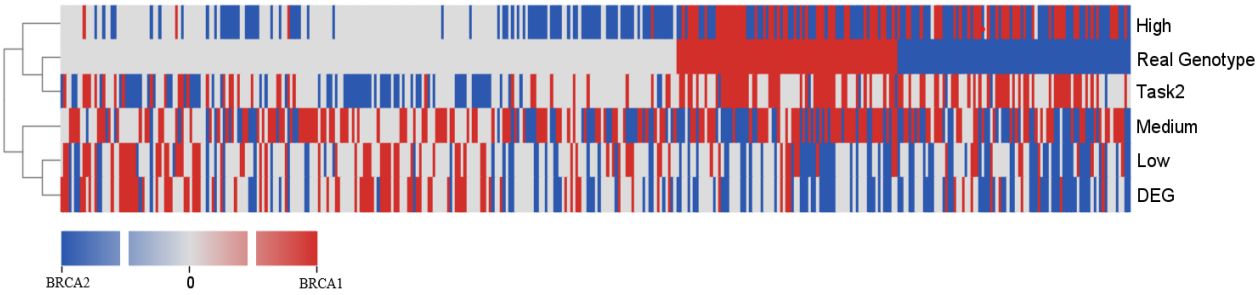


**Figure 3.** Cluster outcome for different miRNA candidate set

Real Genotype for the acutual status of each sample. Low, Medium, High the candidates selected and categorized by RSR model, each with 8, 20, 2 miRNAs. Task 2 the top 30 significant miRNAs performed in the limma model that differs BRCA1 and BRCA2. DEG the GSE226445 selected 19 miRNAs

By K-means cluster, it could be found that the most matched group is high scoring group. While the hamming distance of single miRNA hsa-miR-340-5p used for clustering is 143, the combination of has-miR-140-5p and has-miR-142-3p bivariate coordinate could give the most precise distribution, suggesting that the match is not made by simple addition effect of two high score candidates. Surprisingly, the originally derived differentially expressed miRNAs filtered by limma model performs bad when it really comes to the classification of BRCA genotype, not even good at distinguishing wild-type from the BRCA mutant. Furthermore, the Task 2 significant candidates also failed in distinguishing BRCA1 and BRCA2, estimating that much of the BRCA2 mutations are in the wildtype set, which is the vital problems performed by high score group as well.

## 4. Dicussion

By deploying numerous limma analysis and scoring p values for all 30 miRNA candidates, we get distinctive groups for clustering and discerning the genotype of BRCA. The two miRNAs from the high score have no intersection with the 19 miRNAs, but still managed to get high score in the two tasks, in that they may be sorted out not because of the insignificance in Task 1, but the inconformity of direction in log fold change between the raw data and the batch-adjusted data and the fluctuation of fold change ratio out of the 0.8-1.25 range. This suggests the irrationality of the 3-step filtration method in ~0 + BRCAstatus + Ovarystatus model. Actually, the 0.8-1.25 range filtration is the most rigorous step in the filtration compared to the intersection of significantly expressed miRNAs both in raw data and batch-adjusted data, while such boundary is hard to elucidate, leading to the loss of some potentially good markers for genotype identification.

Limitation of current study has also been noticed. Though the performance of high score group is obviously better than any other group in the BRCA genotype identification, the average performance is still bad. Even the high score group could only give the 58.4% accuracy of the genotype identification. The group has misjudged many wild-type individuals as BRCA2 and still cannot separate BRCA1 and BRCA2 clearly. The two miRNAs in this group can exactly depict a two-dimensional cluster outcome below. From the scatter plot, it could be seen that the BRCA1 and BRCA2 dots are mixed together despite their significant outcome in limma analysis of both tasks, rendering it difficult to cluster no matter what cluster tools, K-means or DBSCAN are to be used since the two are too close. From the view of limma algorithm, such conundrum is possibly because the significance, or the estimator outcome is only mean-dependent as we have deducted in the aforesaid context. The mean-dependent estimator only guarantees the unbiases of the corresponding genotype expression, but not effective enough. The heteroskedasticity of two variables would make them distribute within each other even under the premise that their mean expression is differed significantly. To solve the conundrum, more significant miRNA or other tumor-suppressed genes related BRCAness introduced to raise the cluster dimension may give some help. The current study is fully based upon dataset GSE226445 with only 227 miRNAs to be examined in human body, which

does not include much more factors into account in view of the BRCAness biological mechanisms such as coding genes mutation and BRCA epigenetic silencing in cis-acting elements. The prevalence of promoter methylation may play a role in defective DNA repair system [10]. Hence, much more statistics should be given consideration subjecting to the genotype identification problem.

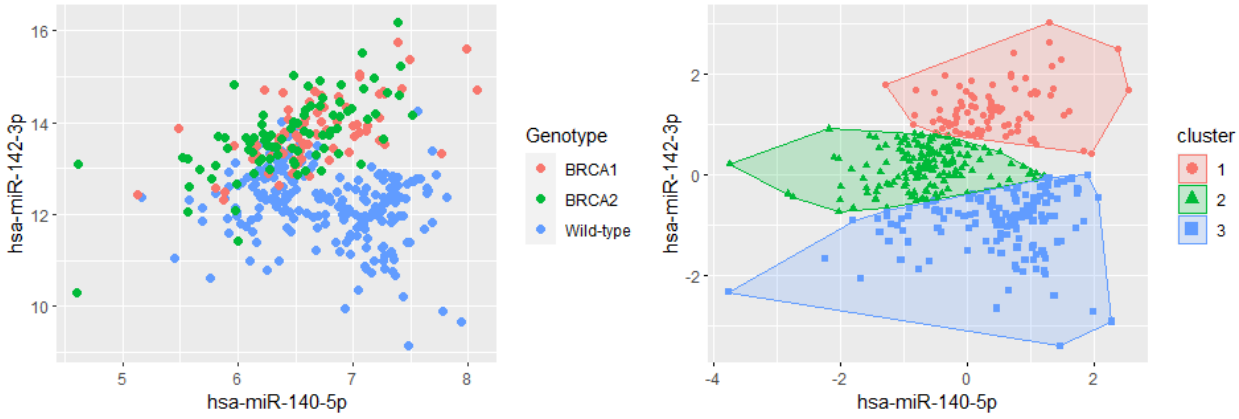


**Figure 4.** K-means cluster of high RSR score miRNA group

**a.** Scatter plot of real genotype for 220+79+83 samples. **b.** K-means visualization of high RSR score miRNA group.

## 5. Conlusion

In conclusion, current study reveals that miRNAs exhibiting differential expression, as discerned through the limma model, may not consistently prove effective in classifying BRCA status, notwithstanding their statistical significance in the respective contrasts. This underscores the need for a deeper exploration of the intricate relationship between genotype identification algorithms and DE gene discovery algorithms, as their congruence is not self-evident in the elucidation of individual clinical data. To enhance the reliability of expression-pattern-based disease diagnosis, the findings advocate for the incorporation of additional bioinformatic approaches, which arises from the recognition that the current paradigm may benefit from a more comprehensive integration of diverse computational methods, thereby refining the precision and robustness of early-stage oncological identification.

## 6. Model Development

### 6.1. K-means clustering

The K-means clustering was performed under R 4.1.0 running under Windows 10 x64 (build 19045), with kmeans function in stats package uniformly set with 250 maximum iteration turns, nstart of 25 and centers of 3 [11]. After giving cluster tags for each subset of tested group, full permutation of {BRCA1, BRCA2, Wild-type} were given to the cluster tags by turns. Tag assignation with the largest Hamming distance to the real genotype was designated the final cluster outcome.

The visualization of K-means cluster in heatmap is completed by TBtools, and that for high score group was realized under fviz_cluster function in R package factoextra [12-13].

### 6.2. RSR classification

Rank sum ratio (RSR) classification was generated by probit model. The RSR rank for certain object in integer mapping form denotes as r, following the univariate regression model as:

$$\widehat{RSR} = \alpha\, \Phi^{-1}\left(\frac{r}{n}\right) + \beta \tag{9}$$

With the last term $\frac{r}{n}$ timing an adjustment factor $\left(1-\frac{1}{4n}\right)$. Objects with same n-fractile $\frac{r}{n}$ would be given the average fractile. The regression model slope has passed F test significantly (P = 3e-16) with determination coefficient $R^2 = 0.91$. The reference of 3 group classification is adopted by the x coordinate -1 and 1 in the standard gaussian distribution, pointing to the classification bound value of $\beta \pm \alpha$.

## References


[1] Jan HJ. Hoeijmakers. Genome maintenance mechanisms for preventing cancer. Nature. 2001, 411 (6835): 366 - 374.

[2] Suzanne T. Witt, Alana Brown, Laura Gravelsinset al. gray matter volume in women with the BRCA mutation with and without ovarian removal. bioRxiv, 2023: 2011 - 2023.

[3] Thien Kieu Thi Phung, Berit Lindum Waltoft, Thomas Munk Laursenet al. Hysterectomy, oophorectomy and risk of dementia: a nationwide historical cohort study. Dementia and geriatric cognitive disorders, 2010, 30 (1): 43 - 50.

[4] W. A. Rocca, J. H. Bower, D. M. Maraganoreet al. Increased risk of cognitive impairment or dementia in women who underwent oophorectomy before menopause. Neurology, 2007, 69 (11): 1074 - 1083.

[5] Heidi D. Nelson, Miranda Pappas, Bernadette Zakheret al. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: a systematic review to update the US Preventive Services Task Force recommendation. Annals of internal medicine, 2014, 160 (4): 255 - 266.

[6] Kevin Elias, Urszula Smyczynska, Konrad Stawiskiet al. Identification of BRCA1/2 mutation female carriers using circulating microRNA profiles. Nature Communications, 2023, 14 (1): 3350.

[7] Gordon K. Smyth. Limma, linear models for microarray data, Springer, 2005.

[8] Charity W. Law, Kathleen Zeglinski, Xueyi Donget al. A guide to creating design matrices for gene expression experiments. F1000Research, 2020, 9.

[9] GFV Glonek, P. J. Solomon. Factorial and time course designs for cDNA microarray experiments. Biostatistics, 2004, 5 (1): 89 - 111.

[10] Weijie Poh, Robert L. Dilley, Alison R. Moliternoet al. BRCA1 promoter methylation is linked to defective homologous recombination repair and elevated miR-155 to disrupt myeloid differentiation in myeloid malignancies. Clinical Cancer Research, 2019, 25 (8): 2513 - 2522.

[11] R. Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2021.

[12] Chengjie Chen, Hao Chen, Yi Zhanget al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Molecular plant, 2020, 13 (8): 1194 - 1202.

[13] Alboukadel Kassambara. Factoextra: extract and visualize the results of multivariate data analyses. R package version, 2016, 1.