Correlation analysis between PIK3CA, TP53, CDH1 genes mutations and breast cancer

Dongyue Zhu, Yunjing Gu, Ping Zhu+
School of Science, Jiangnan University, Wuxi, China
e-mail: zhuping@jiangnan.edu.cn

***Abstract*— In the DNA sequence, accumulation of single nucleotide variation (SNV) changes the amino acid sequence of oncoprotein, and it causes normal cells to turn into cancer cells. This study uses the 1105 samples of breast cancer in the cBioPortal Cancer Database to screen out three key genes, and use the electron-ion interaction pseudopotential (EIIP) of base A, base C, base G and base T to propose the E difference value formula, further verification of the impact of the three key genes on breast cancer patients. The study found that the PIK3CA, TP53 and CDH1 are key genes that cause breast cancer, and PIK3CA, TP53 and CDH1 are closely related to the survival time of breast cancer.**

Keywords-Breast Cancer, EIIP, E difference value

I. INTRODUCTION

Breast cancer is a female high-grade malignant disease, accounting for 23% of all female cancer cases, and leads to 14% of cancer-related death in cases [1]. China used to be a low-incidence country for breast cancer, however, owing to the changes of people's eating habits, reproductive behaviors and lifestyles, the risk of breast cancer is continuously increasing [2]. Data from cancer surveillance sites in Beijing, Shanghai and Harbin have shown that the incidence of breast cancer is rising [3-5]. For the treatment of cancer, early detection, early treatment, the sooner the discovery, the better the treatment. Until now, Studies have shown that synonymous codons have different efficiencies in translational speed and folding accuracy [6]. Single Nucleotide Changes, and it causes different usage of codons, ultimately, the protein's translation efficiency, translation speed and folding accuracy will all change. In study, the DNA sequence is represented digitally by using the Electron-ion interaction pseudopotentials (EIIP) of the bases A, C, G, T. The EIIP value is given by Nair et al [7,8]. Then, this study find out the correlation between genetic variation and EIIP. Gene mutation types include base substitutions, frameshifts, insertions, and deletions. There are two main types of base substitutions: transition and transversion. Defining purine and purine, pyrimidine and pyrimidine mutations is called transition, the mutation between purines and pyrimidines is called transversion. And the ratio of transition and transversion is generally not equal, it is called "conversion bias [9].

II. Materials and Methods

A. Sources of materials

In this study, the data comes from the cBioPortal database(http://www.cbioportal.org), breast invasive carcinoma (TCGA, Provisional , 1105samples).Among 11417 mutated genes in 975 mutated samples, we detected the key genes were significantly associated with breast cancer.

B. Social network algorithm screen out key genes

The concept of social network first came from the social field, mainly refers to the sum of the relationship between one person and others. Social network is mainly to study the sparsity of relational connections. Currently, the mining of relational data [10,11] has become one of the most popular research topics in data mining. Cancer is caused by genetic mutations or some environmental factors. We studies 975 samples of mutations, and up to 3412 genes mutations or at least 1 gene mutations in a sample. In this study, social network was constructed by igraph package in R (https://www.r-project.org),

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00086

searching for important hub genes among a total of 11,417 variant genes in 975 samples. And using the spin-glass social classification function in social network algorithm to analyze the gene network. This study selects data without direction metrics. Then set the background of the image to white and set the node size as follows: if the gene with a center degree greater than 10000, the node size is set to 10. If the gene has a center degree less than 5000, the node size is set to 2. If the center degree is less than 10,000, and greater than 5000 the node size is set to 5. Due to the large number of genes analyzed, we uses the layout function to debug the node's color, node size, and chooses not to display the gene names of each node. Therefore, it can be seen from Figure 1 that PIK3CA, TP53 and CDH1 are the three genes with the highest frequency of breast cancer mutation. Through data review, it has been proved that PIK3CA, TP53 and CDH1 genes play an important role in breast cancer [12-15].

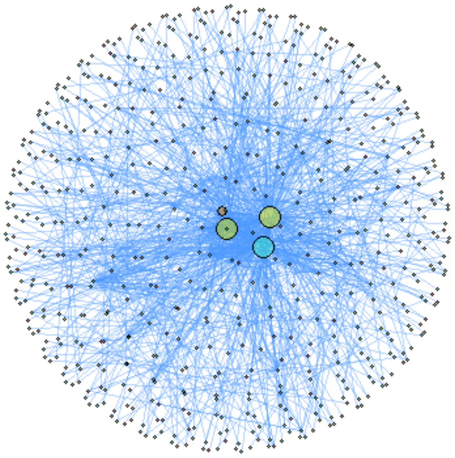

Figure 1. Social network algorithm screen out key genes

C. E difference value

With the development of bioinformatics, there are many ways to numerically map bases in a sequence. Such as Voss mapping, real number mapping, Z-curve mapping, and the mapping established by Zhu Ping[16]: $\varphi: GF(7^3) \rightarrow C_{343}$, et al. Nair and Rao [7,8] proposed that digitizing DNA sequences use nucleotide electron-ion interaction pseudopotential (EIIP) values. The EIIP value of the amino acid sequence has been used in place of the protein's amino acid sequence for resonance recognition model (RRM) extraction information [17]. Digital mapping of bases based on the order in which bases appear in the sequence, the 4 bases EIIP value is shown in TABLE I.

TABLE I. EIIP Table

base	A	C	G	T
EIIP	0.1260	0.1340	0.0806	0.1335

According to the results of the Section *B*, PIK3CA, TP53 and CDH1 have the highest frequency of mutations in breast cancer. This study requires all the samples from TCGA breast invasive carcinoma (TCGA, provisional) with RNA-seq v2 data (n = 1105), and considered RNA dysregulation with Z-score threshold: ±2, and mark HER2 negative or positive, HER2 is an important criterion for judging the severity of breast cancer in medicine. Screening samples by the above conditions. Finally, 416 samples of at least one gene mutation of PIK3CA, TP53 and CDH1 were obtained. And the number of single base A, C, G and T mutation were 108, 92, 165 and 33, respectively. Combined with TABLE I, it can be seen that the EIIP value of T is the highest, whereas the number of T mutations is the lowest, and the EIIP value of G is the lowest, whereas the number of G mutations is the highest. To consider the correlation between base mutation and EIIP value, we propose the following E difference value formula. In order to facilitate the writing of the formula, we defines genetic variation as t, when the base substitution occurs, the base mutation is defined as t=1, other base variants including insertions, deletions, frame shifts, etc, it is defined as t=-1, when no base mutation has occurred, the definition is t=0. E difference value formula is as follows:

$$D_E = \begin{cases} \sum_{i=1}^{n} 2\cdot\left|(E_j - E_i)\right| & t=1 \\ (\mathrm{E}_A+\mathrm{E}_C+\mathrm{E}_G+\mathrm{E}_T)\big/4 & t=-1 \\ 0 & t=0 \end{cases} \qquad (1)$$

n is the number of base substitutions per gene in a sample, E_j is the EIIP value of the wild-type base, E_i is the EIIP value of the base after the substitution. E_A, E_C, E_G, E_T are the EIIP values of bases A, C, G and T.

To further understand the correlation between EIIP values and single base mutations, this study uses E difference value to perform heat map analysis on genetic variation samples. Then, we analyze the mutations of PIK3CA, TP53 and CDH1 in 416 samples, as shown in Figure 2. This study uses the E difference value formula to calculate the EIIP value changes of genes mutation. Then, heat map analysis was performed using the Kendall's Tau matrix and single linkage clustering method by the use of MeV software (https://sourceforge.net/projects/mev-tm4/ files / mev-tm4/). Each small square of the heat map represents a sample, and color indicates the size of the EIIP difference (red is upregulation, green is downregulation). MeV is a gene expression pattern and differentially expressed gene analysis software based on Java application design. In order to further prove the effect of PIK3CA, TP53 and CDH1 genes mutation on breast cancer, in section *C*, MeV software was used to analyze the gene mRNA aberrantly expression of PIK3CA, TP53 and CDH1.

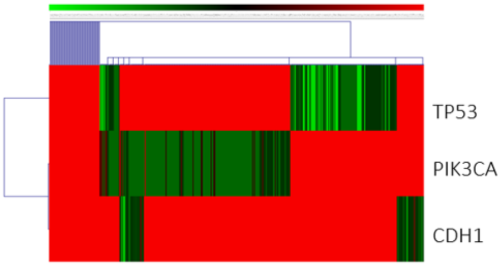

Figure 2. E difference formula to calculate genetic variation (the downregulation value is 0; the upregulation value is 0.3)

D. Method comparison

In MT Birgani's[18] study on the genes mutation of liver cancer, by using mRNA expression level. In this study, we used the Kendall's Tau matrix and single linkage clustering method of MeV software to analyze the genetic mRNA aberrantly expression of PIK3CA, TP53 and CDH1 in breast cancer, See Figure 3. Comparing Fig. 3 with Fig 2: In Figure 2, the samples distribution of each gene mutation is relatively concentrated, and the time used is 48ms. In Figure 3, samples are clustered by mRNA aberrantly expression, and the mutated and non-mutated samples were staggered and the distribution of the mutated samples was relatively messy, and the time used is 65ms. In general, the effect of clustering in Figure 3 is not as good as in Figure 2. From the above analysis we can draw: The E difference values proposed in this study is better when using the Kendall's Tau matrix and single linkage clustering method of MeV software, and the calculation time is relatively short.

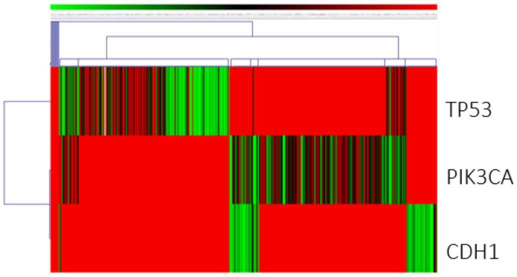

Figure 3. Shows the difference in mRNA expression. (the downregulation value is 0; the upregulation value is 1)

Using the E difference value to analyze the base mutations of breast cancer and the use of mRNA aberrantly expression to analyze the genes mutation of breast cancer are compared, and there are three clustering methods in Table 2. The E difference value has an advantage in the time by the single linkage clustering method. When the data for studying cancer gene mutations are relatively large, it is better to use E differences value to analyze the genes and samples in the single linkage clustering method, which is not only of short time but also of good clustering effect.

TABLE II. MeV software calculation time comparison

method	Current metric	single linkage clustering	average linkage clustering	complete linkage clustering
mRNA expression level	Euclidean distance	38ms	40ms	66ms
	Manhattan distance	44ms	35ms	48ms
	Average dot product	75ms	73ms	85ms
	Covariance value	54ms	51ms	50ms
	Kendall's tau	65ms	63ms	75ms
E difference value	Euclidean distance	51ms	42ms	60ms
	Manhattan distance	40ms	58ms	62ms
	Average dot product	55ms	53ms	101ms
	Covariance value	46ms	63ms	69ms
	Kendall's tau	48ms	63ms	56ms

Using SPSS software, the mRNA expression levels of PIK3CA, TP53, and CDH1 in 416 breast cancer samples were analyzed using a single sample Kolmogorov-Smirnov, with a standard deviation of 0.301. Using the same test method, the E differential values of PIK3CA, TP53 and CDH1 in 416 breast cancer samples were tested, and the standard deviation was 0.051, which was much less than 0.301. E difference value relatively stable and normal distribution. In order to prove the importance of PIK3CA, TP53 and CDH1 genes in breast cancer, we use survival packages in R software to survival analysis. In the process of survival analysis, mainly using Kaplan-Meier estimator to analyze the survival of 975 samples, of which 416 samples of at least one mutation in PIK3CA, TP53 and CDH1 were detected. As you can see from Figure 4, the samples with PIK3CA, TP53 and CDH1 genetic alterations had a poorer survival as compared to those without alterations.

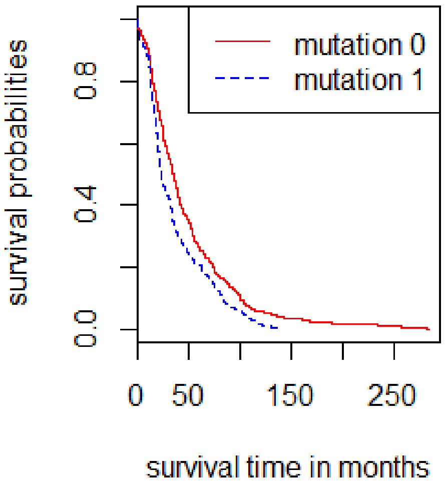

Figure 4. Overall survival analysis. Blue line represents cases with alterations, and red line represents cases without. The X axis indicates overall survival time (months), and the Y axis indicates the survival rate.

E difference value formula for the base mutation analysis has certain help. In MT Birgani's[18] study of genes mutation in liver cancer, mRNA expression level was used to analyze the mutated genes of liver cancer , without considering the mutation of the base. In this study, not only the mRNA aberrantly expression of genes were analyzed, but also an E difference value formula was proposed. From the above analysis, it can be seen that the E difference value has certain advantages both in the heat map clustering and in the single-sample Kolmogorov-Smirnov test.

III. CONCLUSIONS

In the study of breast cancer, general considerations of mRNA aberrantly expression, methylation, copy number variation [18,19], etc, but changes in physical properties of EIIP caused by base mutations are not considered. This study mainly uses 1105 samples from the cBioPortal database, after network analysis, PIK3CA, TP53 and CDH1 were screened out. This study found that the E difference value is better than using the mRNA aberrantly expression for clustering breast cancer samples, especially in the calculation time and clustering effect of the single linkage clustering. This study does a survival analysis of the samples, it was found that PIK3CA, TP53 and CDH1 genes have negative effects on the survival rate of samples. These results suggest that PIK3CA, TP53 and CDH1 genes mutation in the breast cell may offer cancer risk prediction and early detection markers. Overall, the study propounds a potentiality for interpreting the pathogenesis and development of breast cancer with genetic alterations, and provides a novel platform for searching for more capable diagnostic biomarkers for breast cancer.

ACKNOWLEDGMENT

This work supported by the National Natural Science Foundation of China (Grants No. 11271163), Postgraduate Research & Practice Innovation Program of Jiangsu Provence(Grants No. KYCX18 _1865).

REFERENCES

[1] T. Sorlie , CM. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, et al.Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications[J].Proceedings of the national academy of sciences of the united states of America.98(19):10869-10874.

[2] Zhang Q, Liu LY, Wang F, Mu K, Yu ZG. The change in female physical and childbearing characteristic in china and potential association with risk of breast cancer[J]. BMC Public Health,2012,12:368-374.

[3] Song. Bingbing, Wu. Shuling , Han. Huili , Sun. Xiwen, Liu. Wei, Yuan. Weihua, et al. Incidence and mortality of breast cancer in Nangang District of Harbin from 1992 to 2001[J].Chinese Journal of Cancer,2003,12(10):574-576.

[4] Wang. Qijun , Zhu. Weixing, Xing. Xiumei, Li. Ling . Cancer Incidence Trends of Urban Residents in Beijing in 182-1997[J]. Chinese Journal of Cancer,2001,10(9):507-509.

[5] Enju Liu, Yongbing Xiang, Fan Jin, Shuzhen Zhou, Lu Sun, Rurong Fang et al. Analysis of incidence trends of malignant tumors in Shanghai City (1972-1999)[J].Tumor,2004,24(1):11-15.

[6] H. Gingold, Y. Pilpel. Determinants of translation efficiency and accuracy[J]. Molecular systems biology [J]2014,7(1): 481.

[7] AS. Nair, SP. Sreenadhan. A coding measure scheme employing Electron-Ion Interaction Pseudopotential (EIIP)[J].Bioinformation,2006,1(6):197-202.

[8] KD. Rao, MNS. Swamy. Analysis of genomics and proteomics using DSP techniques[J].Circuits & Sytems, 2008,55(1)：370-378.

[9] Zhao. Hui, Li. Qisai, Li. Jun, Zeng. Changqing , Hu. Songnian, Yu. Jun. Study on Neighbor Base Components the Transformation or Transversion of SNPs Generated in Plant Genomes [J]. Chinese Science Series C: Life Sciences, 2006,36(1):1-8.

[10] L. Getoor, CP. Diehl. Link mining :A survey[J]. Acm Sigkdd Explorations Newsletter,2005,7(2):3-12.

[11] Yang . Nan, Gong. DanZhi, Li. Xian, Meng. XiaoFeng. Summary of Web community discovery technology [J]. computer research and development, 2005, 42(3):439-447.

[12] S. Bartels, JL. van Luttikhuizen , M. Christgen, L. Magel, A. Luft, S. Hanzelmann, et al. CDKN2A loss and PIK3CA mutation in myopepithelial-like metaplastic breast cancer[J]. J Pathol, 2018, 245(3):373-383.

[13] SJ. Isakoff, JA. Engelman , HY. Irie , J. Luo, SM. Brachman, RV. Pearline, et al. Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells[J]. Cancer Res,2005,65(23):10992-11000.

[14] G. Watanabe, T. Ishida, A. Furuta, S.Takahashi, M. Watanabe, H. Nakata, et al. Combined immunohis-

tochemistry of PLK1,P21 and p53 for predicting TP53 status: an independent prognostic factor of breast cancer[J]. AmJ Surg Pathol,2015,39(8):1026-1034.

[15] H. Lei, S. Salahshor, B. Werelius, K. Hemminki,A. Lindblom, S. Sjöberg-Margolin , et al. CDH1 mutations are present in both ductal and lobular breast cancer, but promoter allelic variants show no detectable breast cancer risk[J]. Int J Cancer,2002,98:199-204.

[16] Yan YanYan, Zhu Ping. Extended triplet set C343 of DNA sequences and its application to p53 gene[J].Chinese Physic B,2011,20(1):689-697.

[17] J. Cosic, Macromolecular bioactivity: is it resonant interaction between macromolecules? --theory and a pplications [J].Ieee transactions on biomedical engin eering1994,41(12):1101-1114.

[18] MT. Birgani, M. Hajjari, A. Shahrisa, A. Khoshnevisan, Z. Shoja, P. Motahari , et al. Long non-coding RNA SNHG6 as a potential biomarker for hepatocellular carcinoma[J]. Pathology & Oncology Research, 2017(1):1-9.

[19] Gao Y, Widschwendter M, Teschendorff AE, DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants[J]. EBIOMEDICINE,2018,31:243-252.