Analysis of the correlation between key gene mutations and breast cancer

Ziyuan Shen, Dongyue Zhu, Yunjing Gu and Ping Zhu

Abstract
The purpose of this study is to investigate whether mutations in key genes affect the prognosis of breast cancer patients, and base mutations is discussed by combining the physical properties of the electron–ion interaction pseudopotential. We study 994 breast cancer mutation samples, including 15,015 mutation genes, and use social network algorithm to screen key genes. In order to analyze the relationship between electron–ion interaction pseudopotential and mutation of key genes, this paper proposes the $E$ difference formula. It not only analyses gene mutation at the physical level, but also explains the mutation rule of single base. Simultaneously, we use the Kaplan–Meier to analyze the overall survival rate of breast cancer samples and use Tarone-Ware to test. The results showed that mutations in GATA3, PIK3CA, and TP53 are significant difference in overall survival rate of breast cancer ($CI = 95\%$, $P < 0.05$). Proteins often interact to form protein complexes, and driving protein function. Therefore, by constructing protein–protein interaction network and finding modular proteins of key genes, we can find genes closely related to key genes. Finally, the results of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and gene ontology function enrichment analysis show that key genes played an important role in metabolic pathways and biological functions. In clinical medicine, screening for key gene mutations may help predict survival in breast cancer patients.

Keywords
Breast cancer, key gene, EIIP, $E$ difference value, survival analysis

Received 28 January 2019; Revised received 10 May 2019; accepted 16 June 2019

Introduction
Breast cancer is a malignant disease with high incidence in women, accounting for 23% of all female cancer cases, and leads to 14% of cancer-related death in cases. 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. Data from cancer surveillance sites in Beijing, Shanghai, and Harbin have shown that the incidence of breast cancer is rising. The treatment of breast cancer is early detection and early treatment, and the sooner the discovery, the better the treatment. Previous studies have shown that synonymous codons have different efficiency in translation speed and folding accuracy. Changes in single nucleotides lead to different codon usage, that will cause the protein’s translation efficiency, translation speed, and folding accuracy will all change. In the 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 and Sreenadhan, Rao and Swamy. In past genetic mutation studies, gene expression profiles were commonly used, but gene expression profiles did not highlight changes in single base. Therefore, we use the physical properties of the base to calculate the changes of the base after mutation, and this study find out the correlation between base muta-
tions 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 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”.

School of Science, Jiangnan University, China

Corresponding author:
Ping Zhu, School of Science, Jiangnan University, Wuxi, China. Email: zhuping@jiangnan.edu.cn

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Materials and methods

Sources of materials

From the cBioportal database (http://www.cbioportal.org), we downloaded 1105 breast invasive cancer samples, including 994 mutant samples and 15,015 mutant genes.

Social network algorithm screen out key gene mutations

The concept of social network first came from the social field, mainly refers to the sum of the relationship between one person and others, and mainly studies the sparsity of relational connections. Since 1954, anthropologist Barnes first used social network analysis to study social structure. Until 1960, social network analysis was clearly defined as a methodology and developed rapidly in interdisciplinary fields, which was widely used in sociology, anthropology, economics, and psychology.10 Barry,11 a social network analyst, points out that social networks are huge networks of social relationships among groups. Currently, the mining of relational data12,13 has become one of the most popular research topics in data mining. Cancer is caused by gene mutations or some environmental factors. We studied 994 mutation samples, with a maximum of 3412 gene mutations or at least one gene mutation. In this study, the social network was constructed from the igraph package in R (https://www.r-project.org) with a total of 15,015 mutant genes in 994 samples. And using the spin-glass social classification function in social network algorithm to analyze the gene network. After a series of experiments, we find out that when center degree is 4000-8000, PIK3AC, TP53, CDH1 and GATA3 can be screened by the algorithm. When center degree is less than 4000, the screened genes increase gradually, and center degree reflects the importance of this node in the network. Therefore, this paper sets center degree of 4000 as an important segmentation point. And this study selects data without direction metrics, sets the background of the image to white and sets the node size as follows: if the gene with a center degree greater than 4000, the node size is set to 10. If the gene has a center degree less than 4000, the node size is set to 2. Therefore, as can be seen from Figure 1, PIK3CA, TP53, CDH1, and GATA3 are key gene mutations in breast cancer.

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 Yan and Zhu 14: \( \phi : GF(7^3) \to C_{343} \). Nair and Rao7,8 proposed that digitizing DNA sequences use nucleotide EIIP values. The EIIP value of the amino acid sequence has been used in place of the amino acid sequence for resonance recognition model extraction information.15 Digital mapping of bases based on the order in which bases appear in the sequence, the EIIP values of bases are shown in Table 1.

The results of the above analysis show that PIK3CA, TP53, GATA3, and CDH1 are key gene mutations in breast cancer. Among the key gene mutations, the number of substitution mutations of single base A, C, G, and T was 213, 169, 252, and 71, respectively. Combined with Table 1, 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 mutations and EIIP value, we propose the following \( E \) difference value formula. In order to facilitate the writing of the formula, we define gene mutations as \( t \). When the base substitution occurs, the base mutation is defined as \( t = 1 \), other base mutations including insertions, deletions, frame shifts, and so on, it is defined as \( t = -1 \). When no mutation occurs, we define \( t = 0 \). \( E \) difference value formula is as follows:

\[
D_E = \begin{cases} 
\sum_{i=1}^{n} 2 \cdot |(E_j - E_i)| & t = 1 \\
(E_A + E_C + E_G + E_T)/4 & t = -1 \\
0 & t = 0
\end{cases}
\]  

(1)

![Figure 1. Social network algorithm screen out key gene mutations.](image)
\( n \) denotes the number of base substitutions in a gene, \( E_i \) 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 breast samples. Then, we analyze the mutations of GATA3, CDH1, PIK3CA, and TP53 in 1105 samples, as shown in Figure 2(a). This study uses the \( E \) difference value formula to calculate the EIIP value changes of gene mutations. 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 value (red is high \( E \) difference value, green is low \( E \) difference value).

**Result**

**Differential gene expression**

In MT Birgani’s study on the genes mutation of liver cancer, by using mRNA expression level. In this study, we use Kendall’s Tau matrix and single linkage clustering methods to analyze differential expression of key genes in breast cancer samples, as shown in Figure 2(b). According to this figure, GATA3, CDH1 are highly expressed in breast cancer samples, PIK3CA, TP53 are low expressed in breast cancer samples. In breast cancer samples, Figure 2(a) mainly reflects the mutation of key genes. Figure 2(b) mainly reflects the expression of genes, which cannot reflect gene mutations. And in the study of gene mutation, \( E \) difference value can better reflect the mutation of gene, and the clustering effect of mutation samples is better. At the same time, when clustering by Euclidean distance, Manhattan distance, average dot product, covariance value, and other methods in MEV software, the \( E \) difference value clustering effect is better, and it also reflects the mutation of genes in samples.

**Survival analysis**

To analyse the importance of PIK3CA, TP53, CDH1, and GATA3 in breast cancer, we used the Kaplan–Meier method for survival analysis and tested it with the Tarone-Ware method. As shown in Figure 3(a) to (c) where 0 indicates no mutation in key genes, 1 represents at least one gene mutation in key genes. After survival analysis, we found that TP53, GATA3-TP53, GATA3-PIK3CA-TP53 mutations have significant differences in survival rates for breast cancer patient, and \( P \) values were 0.022, 0.041, 0.04, respectively. Based on the above analysis, we conclude that there is a significant difference in survival rate of breast cancer patients with key gene mutations. And it has certain clinical significance for the prognosis of breast cancer patients.

Proteins often interact to form protein complexes, and driving protein function. So through the analysis of protein–protein interaction (PPI) network, key proteins can be found. Using PIK3CA, TP53, GATA3, and CDH1 as seed genes, the interaction genes were queried by cBioPortal database. PPI network in Figure 3(d) from cBioPortal database (http://www.cbiportal.org/). Then this paper use MCODE plug-in to modularize PPI, as shown in Figure 3(e). This analysis primarily helps us discover genes that are more closely linked in the PPI network. However, no genes were found to be more closely associated with GATA3.

**KEGG pathway and GO enrichment analysis**

In order to study the functional mechanism of 48 genes with the largest number of nodes (nodes$>415$) in social networks, KEGG pathway analysis and gene ontology (GO) enrichment analysis are made by using DAVIA (https://david.ncifcrf.gov/) online website. The 48 genes were significantly enriched (\( P < 0.05 \)) in 56 KEGG pathways, and the 20 paths with the lowest \( P \) value are analyzed and shown in Figure 4. The 48 genes were mainly enriched in HTLV-I infection, pathways in cancer, colorectal cancer, hepatitis B, endometrial cancer, chronic myeloid leukemia, MicroRNAs in cancer, pancreatic cancer, melanoma, thyroid hormone signaling pathway, and so on. In particular, PIK3CA enriched 18, and TP53 enriched 17, in the first 20 pathways, so it can be seen that PIK3CA and TP53 are important genes not only in breast cancer, but also in other cancers. Gene enrichment in cancer-related

---

**Figure 2.** Head map. (a) \( E \) difference formula to calculate genetic mutation. (b) Differential gene expression.
The biological significance of the 48 genes was analyzed using GO annotations at three levels: biological process, molecular functions, and cellular component. The 48 genes were significantly enriched ($P < 0.01$) in 32 terms, and as shown in Figure 5. The 48 genes were highly enriched in positive regulation of transcription from RNA polymerase II promoter, negative regulation of transcription from RNA polymerase II promoter, and in utero embryonic development in biological processes. These genes were significantly enriched in ATP binding in the molecular functions. These genes were significantly enriched in nucleus and nucleoplasm in the cellular component. The distribution of key genes in GO enrichment is shown in Table 3. It can be seen from the table that GATA3 and TP53 genes are involved in multiple biological processes, and are in multiple cells to exert multiple molecular functions. It can be seen that GATA3 and TP53 have important positions in the breast and other

Figure 3. Survival analysis and protein–protein interaction network (PPI). (a) TP53 mutations survival analysis, (b) GATA3-TP53 mutations survival analysis, (c) GATA3-PIK3CA-TP53 mutations survival analysis, (d) PPI network, (e) modularize PPI.
parts of the human body. The PIK3CA gene is mainly enriched in the phosphatidylinositol 3-kinase complex, in the cell composition. So it plays an important role in regulating cell proliferation, differentiation, survival, and migration. CDH1 gene is mainly enriched in positive regulation of transcription and DNA-templated, in biological processes. So this gene is mainly involved in the transcription process of cells.

**Conclusion**

PIK3CA is an important signal transducer involved in cell proliferation, a large number of clinical studies and practical experience show that: patients with specific mutations of PIK3CA gene are resistant to EGFR and HER2 targeted drug therapy. In breast cancer patients, the mutation frequency of PIK3CA is as high as 30%, mutation of the gene may lead to disorder or activation of signaling pathway.\(^{17}\)

P53 is a widely studied tumor suppressor gene, when mutated into TP53, the space conformation of P53 has changed, and its anti-cancer effect has also changed.\(^{18}\) According to the latest data of TP53 mutation database WHO IARC TP53 http://www-p53.iarc.fr/, TP53 mutation rate was the highest in tonsillar and female genital tumors.\(^{19}\)

CDH1 (E-cadherin) is an important member of the cadherin family. Its abnormal expression reduces cell adhesion, makes cells easy to separate from surrounding tissues and is

---

**Figure 4.** KEGG pathway analysis.

**Table 2.** Enriched KEGG pathways and the corresponding target genes.

| KEGG pathway                  | Count | P-value           | Target genes                                                                 |
|-------------------------------|-------|------------------|------------------------------------------------------------------------------|
| Colorectal cancer             | 8     | 7.67 × 10\(^{-9}\) | MSH6, MSH2, SMAD4, TP53, PIK3CA, PIK3R3, PIK3R1, APC                        |
| Endometrial cancer            | 7     | 1.11 × 10\(^{-7}\) | TP53, PIK3CA, CDH1, PIK3R3, PTEN, PIK3R1, APC                                |
| Chronic myeloid leukemia      | 7     | 7.11 × 10\(^{-7}\) | SMAD4, TP53, PIK3CA, PIK3R3, RUNX1, ABL1, PIK3R1                             |
| Pancreatic cancer             | 6     | 9.19 × 10\(^{-6}\) | SMAD4, TP53, BRCA2, PIK3CA, PIK3R3, PIK3R1                                  |
| Melanoma                      | 6     | 1.53 × 10\(^{-5}\) | TP53, PIK3CA, CDH1, PIK3R3, PTEN, PIK3R1                                   |
| Glioma                        | 5     | 2.03 × 10\(^{-4}\) | TP53, PIK3CA, PIK3R3, PTEN, PIK3R1                                         |
| Small cell lung cancer        | 5     | 5.19 × 10\(^{-4}\) | TP53, PIK3CA, PIK3R3, PTEN, PIK3R1                                         |
| Prostate cancer               | 5     | 6.21 × 10\(^{-4}\) | TP53, PIK3CA, PIK3R3, PTEN, PIK3R1                                         |
closely related to tumor protein metastasis. Some studies have found that CDH1 gene mutations are found in hereditary diffuse gastric cancer, ovarian cancer. GATA3 is one of GATA transcription factor families, also known as GATA binding protein 3, and it has an important function to form and maintain the differentiation of hemoglobin cells. Studies have shown that GATA3 is highly expressed in breast cancer tissues and is associated with the occurrence, development, and prognosis of breast cancer. In the study of breast cancer, general considerations of gene expression, methylation, copy number variation, but changes in physical properties of EIIP caused by base mutations are not considered. In this study, key genes were screened from a large number of mutant genes in breast cancer by using social network algorithm. Most importantly, when studying gene mutation in cancer, difference is better than using gene expression. In this paper, the survival analysis was carried out, and it was found that key gene mutations were significantly associated with the prognosis of breast cancer. Simultaneously, PPI analysis means finding genes that are closely associated with key gene mutations. The functional mechanisms of the 48 genes were analyzed by KEGG pathway enrichment and GO enrichment. By enrichment analysis, we found that PIK3CA, TP53, GATA3, and CDH1 mutations 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 method for searching for more capable diagnostic biomarkers for breast cancer.

Acknowledgments
This work supported by the National Natural Science Foundation of China (Grant No. 11271163), Postgraduate Research & Practice
Innovation Program of Jiangsu Provence (Grant No. KYCX18_1865).

**Declaration of conflicting interests**
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**
The authors received no financial support for the research, authorship, and/or publication of this article.

**Endnote**
This is a revised version of a paper, which named the Correlation analysis between PIK3CA, TP53, CDH1 genes mutations and breast cancer, published in the proceeding of the DCABES 2018 Conference. Special collection selected as Big Data and its Applications.

**ORCID iDs**
Ziyuan Shen https://orcid.org/0000-0001-9227-1673
Dongyue Zhu https://orcid.org/0000-0001-7457-5104

**References**
1. Sortie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98: 10869–10874.
2. Zhang Q, Liu LY, Wang F, et al. The change in female physical and childbearing characteristic in china and potential association with risk of breast cancer. BMC Public Health 2012; 12: 368–374.
3. Song BB, Wu SL, Han Hl, et al. Incidence and mortality of breast cancer in Nangang District of Harbin from 1992 to 2001. Chinese J Cancer 2003; 12: 574–576.
4. Wang QJ, Zhu WX, Xing XM, et al. Cancer incidence trends of urban residents in Beijing in 182-1997. Chinese Journal of Cancer 2001; 10: 507–509.
5. Liu EJ, Xiang YB, Jin F, et al. Analysis of incidence trends of malignant tumors in Shanghai city (1972–1999). Tumori 2004; 24: 11–15.
6. Gingold H and Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol 2014; 7: 481.
7. Nair AS and Sreenadhan SP. A coding measure scheme employing Electron-Ion Interaction Pseudopotential (EIIP). Bioinformation 2006; 1: 197–202.
8. Rao KD and Swamy MNS. Analysis of genomics and proteomics using DSP techniques. Circuits Syst 2008; 55(1): 370–378.
9. Zhao H, Li QS, Li J, et al. Study on neighbor base components the transformation or transversion of SNPs generated in plant genomes. Chinese Sci Ser C: Life Sci 2006; 36: 1–8.
10. Liu SY, Shi YF, Liu Z, et al. Research on the application of collective interactive learning analytics under the web environment-form the perspective of social network analysis. China’s Electrif Educ 2017; 6: 114–119.
11. Barry W. Network analysis: some basic principles. Social Theory 1983; 1: 155–200.
12. Diehl CP. Link mining: A survey. ACM SIGKDD Explor Newslett 2005; 7: 3–12.
13. Yang N, Gong DZ, Li X, et al. Summary of web community discovery technology. Comput Res Dev 2005; 42: 439–447.
14. Yan YY and Zhu P. Extended triplet set C343 of DNA sequences and its application to p53 gene. Chinese Phys B 2011; 20: 689–697.
15. Cosic J. Macromolecular bioactivity: is it resonant interaction between macromolecules–theory and applications. IEEE Trans Biomed Eng 1994; 41: 1101–1114.
16. Birgani MT, Hajjari M, Shahrisa A, et al. Long non-coding RNA SNHG6 as a potential biomarker for hepatocellular carcinoma. Pathol Oncol Res 2018; 24(2): 329–337.
17. Lambert A, Salleron J, Lion M, et al. Comparison of three real-time PCR assays for the detection of PIK3CA somatic mutations in formalin-fixed paraffin embedded tissues of patients with breast carcinomas. Pathol Oncol Res: POR 2018; 25(3): 1117–1123.
18. Gazdar AF, Bunn PA and Minna JD. Small-cell lung cancer: what we know, what we need to know and the path forward. Nat Rev Cancer 2017; 17: 725–737.
19. Olivier M, Eeles R, Hollstein M, et al. The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 2002; 19: 607–614.
20. Tan YC, Jia YM and Lei KJ. CDH1 gene and ovarian cancer. J Chuambei Medical College 2015; 30: 896–901.
21. Shen ZS, Zhou CC, Li JY, et al. The association, clinicopathological significance, and diagnostic value of CDH1 promoter methylation in head and neck squamous cell carcinoma: a meta-analysis of 23 studies. Onco Targets Ther 2016; 9: 6763–6773.
22. Cardoso MFS, Castelletti CHM, Lima-Filho JL, et al. Putative biomarkers for cervical: SNVs, methylation and expression profiles. Mutat Res 2017; 773: 161–173.
23. Kaurah P, Mac Millan A, Boyd N, et al. Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. JAMA 2007; 297: 2360–2372.
24. Jiang LL, Xin XY, Zhou J, et al. Effect of CDH1 gene mutation/methylation on the expression of cadherin in epithelial ovarian cancer. Journal of Shanxi Medical University 2010; 41(3): 214–218.
25. Cheng Kai, Yu Bo, Wu Nan, et al. Research progress of GATA3 in tumor diagnosis. Chinese J Diagn Pathol 2015; 22: 439–442.
26. Clark BZ, Beriwal S, Dabbs DJ, et al. Semiquantitative GATA3 immunoreactivity in breast, bladder, gynecologic tract, and other cytotkeratin 7-positive carcinomas. Am J Clin Pathol 2014; 142: 64–71.
27. Liu H, Shi J, Wilkerson ML, et al. Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: a useful immunomarker for breast and urothelial carcinomas. Am J Clin Pathol 2012; 138: 57–64.
28. Gao Y, Widschwendter M and Teschendorff AE. DNA methylation patterns in normal tissue correlate more strongly with breast cancer status than copy-number variants. EBioMedicine 2018; 31: 243–252.