Increase the Accuracy of Detection of Pathogenic Genes of Breast Cancer using a Graph-Based Approach to the Gene Prioritization Problem

Mohammed Thajeel Abdullah
Karbala Technical Institute, Al-Furat Al-Awsat Technical University, Kerbala, Iraq
inkr.moh4@atu.edu.iq

Abstract—Cancer is one of the most common causes of mortality today. This disease's complications impose many costs on the human community's health, care, and well-being sectors. Solving complex biological problems requires advanced computational methods, and bioinformatics was created to solve such complex problems with the active interaction of several fields of science. Bioinformatics is an interdisciplinary science combining biological sciences, computers, mathematics, and statistics. The issue investigated in this research deals with one of the challenging issues in bioinformatics: candidate gene prioritization in breast cancer. Gene prioritization means sorting genes based on their relevance to a specific disease, such as breast cancer. Finally, the genes are checked according to their importance in costly experiments. The proposed approach in this research is to present a method based on graph mining for prioritizing genes. The study conducted with ENDEAVOUR and DIR methods was compared and evaluated. The evaluation results show that the designed method is more efficient than other methods.

Keywords—Breast Cancer Diagnosis; Increase the Accuracy; Detection of Pathogenic Genes

1 Introduction

A broad spectrum of electric machines is widely used. Cancer is one of the most common causes of mortality today. This disease's complications impose many costs on the human community's health, care, and well-being sectors. Therefore, always finding effective and efficient methods for early diagnosis and rapid and correct treatment of this disease is a significant concern has been a researcher.

The term "cancer" is used for more than 100 different illnesses, including malignant tumors of various regions of the body such as breast, cervical, prostate, stomach, colon, rectum, lung, mouth, leukemia, bone marrow, Hodgkin's disease, and non-Hodgkin's lymphoma, and what in all these diseases are common, a defect in the mechanisms regulating natural cell growth, cell proliferation, and death. In the meantime, breast cancer is the most common cancer and the most common cause of death due to cancer among women. All tumors are not cancerous and may be benign or malignant.

Benign tumors are abnormal but rarely fatal. On the other hand, malignant tumors are newer and cancerous. Early breast cancer diagnosis dramatically reduces women's mortality rate [1].

Two sets of genes are used to solve the problem of ranking candidate genes: the set of candidate genes and the set of known genes in the disease. In the set of known genes for a disease, some genes have already been proven to be linked to that disease. In the set of potential genes, there are unknown genes that we want to find based on how similar they are to the known genes to put the known genes for the disease in order.

Researchers can study the output of the gene prioritization problem for final confirmation in the laboratory. The general strategy of computational methods is to measure the similarity of Candida genes with disease genes, called gene ranking [2].

Conventional genetic studies of breast cancer often do not identify the exact location of the pathogenic genes but rather identify areas within the genome that contain large numbers of Candida genes in breast cancer. Regarding prioritizing candidate genes in breast cancer, candidate genes are computationally prioritized in order of importance. Research in this field can lead to faster detection of breast cancer genes and ultimately be effective in more accurate diagnosis and treatment of this disease [3].

Identification of pathogenic genes often begins with conventional genetic studies. Current laboratory methods cannot determine the exact location of pathogenic genes on the genome. These methods can only identify areas of suspected disease on the genome, which contain many candidate genes. Laboratory evaluation of all Candidate genes in these areas requires exorbitant financial and time costs. Therefore, computational methods to prioritize candidate genes in these areas before using laboratory methods can dramatically change the pathogenic gene detection process.

Identifying the true genes of breast cancer from a large number of candidate genes in the laboratory is very time-consuming and costly, so computational prediction of candidate genes before laboratory analysis is necessary because it saves time and time.

2 Related Works

Various studies have been conducted on the diagnosis of breast cancer. All of these researches have also yielded different results. Regarding computational algorithms, candidate gene prioritization tools are mainly divided into complex network-based methods and similarity-based methods.

There is a general rule in all methods based on complex networks: genes with high and close interactions on the graph can be involved in the same disease [4, 5]. Existing ways based on complex networks are different in defining distance criteria. These methods are usually divided into two categories: local methods and global methods. Local methods are based on direct interactions between proteins or the shortest distance [4-6]. Global methods earn similarity points based on the number of visits to each node on the network. Studies show that global random step algorithms with higher restart and republishing are more efficient than local network-based methods [7, 8].

There are three types of distance criteria in network-based methods for finding disease-related genes:

2.1 Distance Direct neighbor

In this criterion, genes directly between them have a score of one; otherwise, they have an infinite score. It is clear that the use of this criterion, the direct neighbor, is vulnerable to the lost interactions and false positive interactions that abound in protein interaction networks.

Another disadvantage of this criterion is that it does not consider indirect interactions between proteins. Because, as shown in previous work such as [9], proteins that do not interact directly but have common neighbors or are close to each other tend to have the same biological functions. Usually, they participate in the same biological pathways [10]. Some methods, such as [11], count the number of common neighbors. In other words, candidate genes with more neighbors with known genes in the disease may play a role in the development of the disease.

2.2 The shortest route length criterion

The shortest path length between two biomolecules in a network of molecular interactions is related to the speed of information communication or the degree of functional dependence between the two molecules. Therefore, the shortest path length is a good criterion for expressing the functional similarity between the two genes [12]. Although this criterion is better than the direct neighbor criterion, not considering alternative ways is one of the drawbacks of this method.

As stated in work [13]. The presence of several pathways between two proteins in the network means a stronger functional relationship between the two proteins, and the strength and resistance of such networks to mutations are improved.

The problem with the shortest path algorithm is that it considers only one of the shortest paths, and how many paths with the shortest path lengths between two proteins are ignored, nor do the paths with larger lengths between the two proteins. There is often more than one route between runs and even more than one shortest route. As mentioned, such pathways indicate a significant relationship between the two proteins.

Another disadvantage of the shortest path algorithm is the low resolution. Because the path lengths in the number grid are integers, and the length of the largest path in biological networks is usually very short, this is due to the small world nature of these grids [14, 15].

Instead of these network-based local methods, global methods that use the entire network topology and multiple paths in the network are more efficient than local methods.

2.3 Global distance

Global distance methods do not have the disadvantages of local methods because they consider the overall network topology. Candida genes are first mapped with all known disease genes in these methods. These methods then map a score to each candidate gene. The score of each candidate gene depends on the location of the gene. The closer the gene is to the disease-related genes, the higher the score of this gene, with a space size of 2.02 times the character size. The text must be fully justified.

3 Proposed Method

In this section, the biological data sources used in this research are first introduced. The general structure of the proposed method is introduced, and it ends with the details of the algorithm used and how to implement it. This research uses Protein-Protein Interaction Network, a human genome dataset, and genomic information in the UCSC database. This set has been used to create a set of breast cancer candidate genes in the method evaluation process. For each known gene in each disease, the first 99 genes from the chromosomal neighborhood of that gene are extracted from this data set, and a set with one hundred candidate genes is obtained. This dataset was downloaded from (http://genome.ucsc.edu/).

Protein interaction networks are one of gene prioritization tools' most widely used data sources. The logic of using these data sources for gene prioritization is that proteins related to a specific disease tend to have many connections in interactive networks. From the point of view of biology, the interaction between proteins occurs when two or more proteins are connected to perform a specific biological function. Many current molecular processes in cells, such as DNA replication, can be considered the result of these protein bonds. So, in general, the interaction network between proteins can be defined as a complex system of proteins connected by interaction.

The results of this research were evaluated by the Leave One Out Cross Validation (LOOCV) method, as most of the existing tools have used this method to measure their accuracy. This method is a special case of the K-Fold CV method when K=1. The LOOCV method removes a known breast cancer gene from the total number of known breast cancer genes, which we call the target gene at each stage. Next, 99 genes in the chromosomal neighborhood of this gene are extracted using the UCSC genomic database, and the collection of these hundred genes forms the list of candidate genes for prioritization, see Fig. 1.

The remaining genes known in breast cancer, i.e. (N-1) other genes, form the genes known in breast cancer. Next, the relationship score between genes and breast cancer is calculated for all the genes in the candidate set. Then the list of candidate genes is sorted based on the scores obtained. After finishing the prioritization, we evaluate the rank of the target gene by searching the prioritized list.

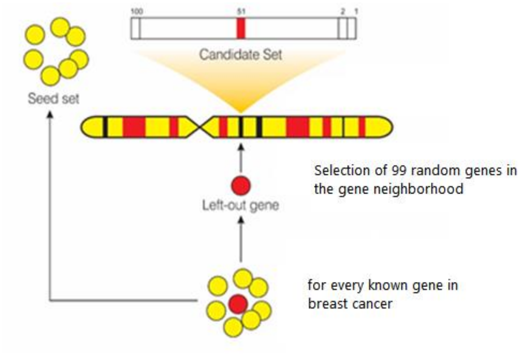
100
51
2
1
Candidate Set
Seed set
Selection of 99 random genes in the gene neighborhood
Left-out gene
for every known gene in breast cancer

Fig. 1. How to generate candidate genes for each target gene in breast cancer.

The best rank equals one, and the worst possible rank is 100. This process is performed for all 25 known genes in breast cancer. Next, the method's efficiency will be calculated by calculating AUC (see Fig. 2).

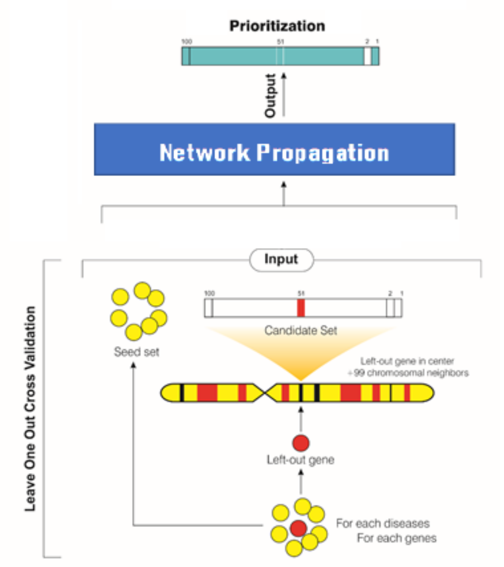

Fig. 2. How to perform the cross-evaluation of the removal of a sample in the GPS tool.

Since only one gene related to breast cancer is included among 99 random genes in the candidate genes, the error rate cannot be a suitable criterion to show the method's efficiency. When the positive and negative samples in a data set do not have a proportional distribution, one of the appropriate ways to calculate the efficiency is to use the area under the ROC diagram.

A proportion of the genes involved in breast cancer located below the threshold indicates a True Positive Rate (TPR), and a proportion of non-pathogenic genes located below the threshold shows a False Positive Rate (FPR). The correct positive rate versus the false positive rate is plotted as a Receiver Operating Characteristic (ROC) curve graph. The area under the ROC chart is called the Area Under Curve (AUC), one of the efficiency measurement criteria. An ideal gene prioritization tool has an AUC value of 1, while a random gene prioritization tool has an AUC value of 5. will be. Therefore, the higher the AUC, the higher the accuracy.

The code used to determine the area under the graph in each prioritized list is in Appendix A.

In addition to the AUC criterion, other criteria, such as TOP1%, TOP5%, and TOP30%, were used to evaluate the methods. These criteria express the number of target genes ranked under one, five, and thirty.

4 Experimental Results

Ranking results based on biological complex networks:

Three network-based algorithms (network propagation algorithm, random step algorithm with restart, and shortest path algorithm) were tested. The evaluation results on the HIPPIE protein interaction network showed that global network-based methods (network diffusion algorithm (84%) and random step algorithm with restart (83%) compared to the local shortest path method (67%) have better accuracy. As seen in Fig 3 and Table 1, both methods based on the global structure of the network have almost the same efficiency. Still, with a slight difference, it is preferable to the network diffusion algorithm.

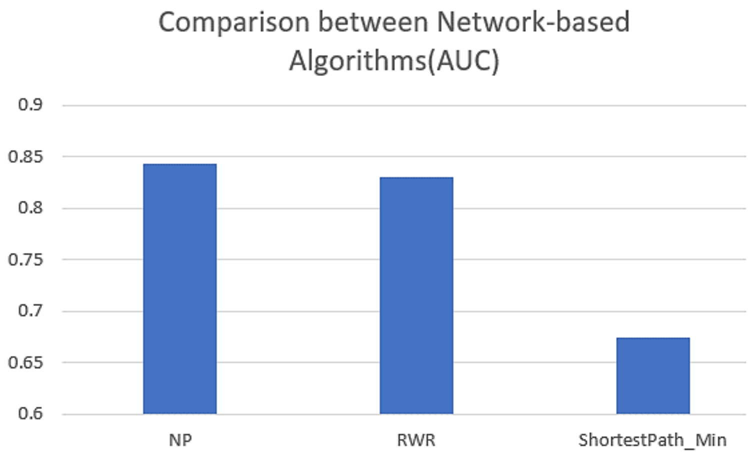

Fig. 3. Comparison of the efficiency of network-based algorithms in ranking breast cancer genes.

Table 1. Evaluation results of network-based algorithms on the HIPPIE data source

Method	**AUC**
Network propagation	0.84
Random walk with restarts	0.83
Shortest distances	0.67

According to the evaluations carried out in the HIPPIE-NP local ranking, the network propagation algorithm was used on the HIPPIE dataset. The network propagation algorithm includes a free parameter called the restart rate alpha (α), which the user should set. In this research, all alpha values are in the range of 1/. Up to 9/. They were checked for network diffusion algorithms. The AUC values for the alpha parameter are shown in Fig. 4 to Fig. 6, and the optimal value for alpha is one-tenth in this study. This evaluation was done on breast cancer with 25 genes.

The method presented in this research was also compared with other tools in this section. Existing tools for prioritizing candidate genes receive different information from the user as input.

Therefore, direct comparison between many tools is impossible or difficult because they receive other inputs. Based on the research on eight famous tools in prioritizing candidate genes, the ENDEAVOUR software achieved a good performance [16]. These eight tools are [17-24].

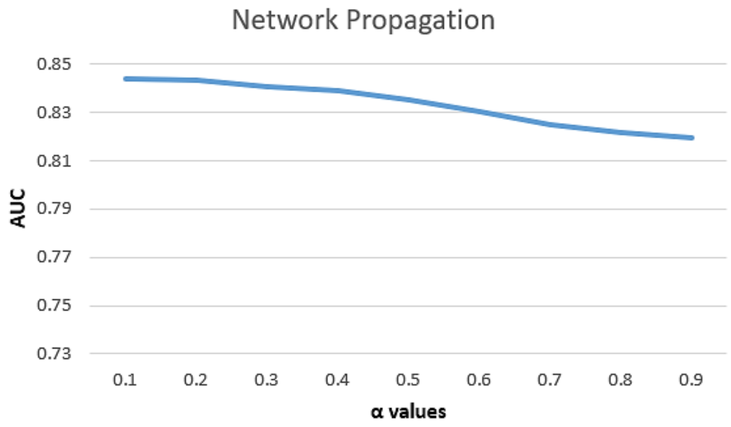

Fig. 4. The effect of different values of the alpha parameter in the range of 0.1 to 0.9 on the efficiency of the network propagation algorithm.

According to the results of this research, ENDEAVOUR software was selected for comparison in this research. In addition, DIR software was also chosen for comparison because this software is similar to our method in terms of the number of resources used.

The results of evaluations on our proposed method and two ENDEAVOUR DIR software are given in Fig. 5. The rank obtained for each gene is also given.

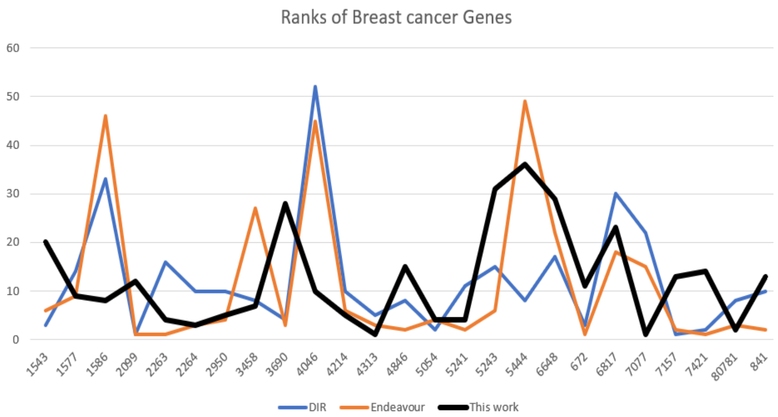

Fig. 5. Comparing the efficiency of the software in the rank obtained for each of the breast cancer genes.

The evaluation results show that our proposed method performs better than other methods in most criteria. In the evaluations, the highest AUC value is related to this research method, and the lowest AUC value is connected to the DIR method Fig. 6.

From the point of view of the number of genes correctly ranked first (TOP1%), our method and the DIR method have almost the same performance, and in both methods, two genes were ranked first. While this criterion in Endeavor software is higher than the above two methods, four genes were ranked first. In all three compared methods, the number of genes placed above the threshold of 30% (TOP30%) is almost close to each other, and more than 80% of genes in all three methods were placed at less than 30.

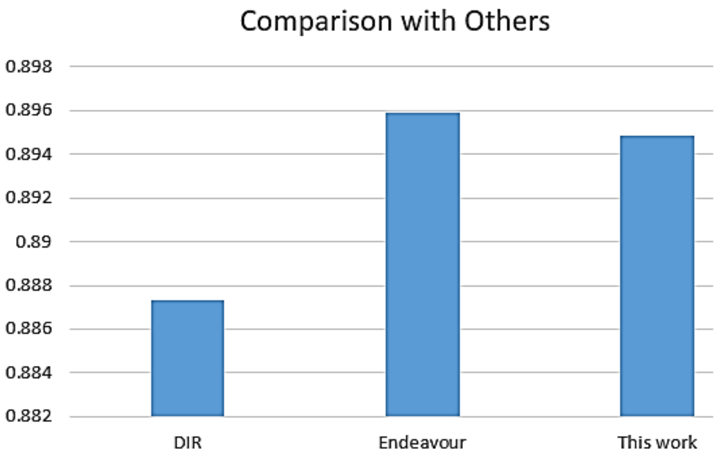

Fig. 6. AUC rate for the three compared tools.

In this section, the method of research evaluation and efficiency criteria were introduced, and then the research results were presented. Finally, the proposed method was compared with two gene prioritization tools.

5 Discussion

In this research, network propagation (NP), random walk with restart (RWR) and shortest path algorithms were implemented. Finally, the post-propagation network algorithm was used in this ranking due to the appropriate efficiency. AUC values for each method are given in parentheses. The results of NP and RWR algorithms are somewhat similar, but the NP algorithm (84%) is marginally better than the RWR algorithm (83%). The evaluations show that NP and RWR national methods outperform the local shortest path method (62%). Because the methods based on the global structure of the network consider most of the interactions in the protein interaction networks in their calculations, this research shows that in the problem of gene prioritization, the network backpropagation algorithm performs slightly better than other global methods. The α parameter makes the algorithm flexible. The values of this parameter indicate the importance of prior knowledge (genes known in breast cancer) versus the importance of the network structure. In this research, the alpha value was set to 10th and used in this ranking. This value for the alpha parameter indicates that more importance is given to the structure of the HIPPIE network.

There are three types of distance criteria in network-based methods for finding genes associated with breast cancer:

Direct neighbor:

In this criterion, genes that have ridges directly between them have a score of one and otherwise have an infinite score. Clearly, this criterion, i.e., direct neighbor, is vulnerable to missing interactions and false positive interactions that are abundant in protein interaction networks.

Another disadvantage of this criterion is not considering indirect interactions between proteins. Because proteins that do not interact directly with each other but have common neighbors or are close to each other in the network tend to have the same biological functions and usually participate in the same biological pathways, some methods, like [10], count the number of familiar neighbors. In other words, candidate genes with more neighbors with known genes in the disease may play a role in the occurrence of the disease.

The length criterion of the shortest path:

The shortest path length between two biomolecules in the network of molecular interactions is related to the speed of information communication or the degree of functional dependence between two molecules. Therefore, the length of the shortest path is a good measure to express the degree of functional similarity between two genes. Although this criterion is better than the direct neighbor criterion, not considering alternative ways is one of the drawbacks of this method.

The presence of several paths between two proteins in the network means a stronger functional connection between the two proteins, and the strength and resistance of such networks against mutations are improved. The problem of the shortest path algorithm is to consider only one of the shortest paths, and how many paths with the shortest path exist between two proteins are ignored, and it also does not consider the paths with a longer length between two proteins. Most of the time, there is more than one path between runs and even more than one shortest path. As mentioned, such pathways indicate a significant connection between two proteins.

Methods based on global distance:

It does not have the disadvantages of local methods. Because these methods consider the overall topology of the network, in these methods, candidate genes are first mapped along with all known breast cancer genes in the network. These methods then map a score to each candidate gene. The score of each candidate gene depends on the location of this gene. The closer the location of the gene is to the genes associated with breast cancer, the higher the score of this gene will be.

6 Conclusion and Future Work

Laboratory diagnosis of pathogenic genes is one of the hundreds of candida genes in breast cancer that is very time-consuming and costly. The enormous number and volume of connections between genes and diseases make studying genetic diseases in humans difficult and even impossible in the laboratory. The vast amount of biological information expanding daily has made using computational methods inevitable. This research could be an effective step in accelerating the process of detecting and detecting genes in breast cancer using computational algorithms.

Identification of pathogenic genes often begins with conventional genetic studies. Current laboratory methods cannot determine the exact location of pathogenic genes on the genome, and these methods can only identify areas of suspected disease on the genome that contain many candidate genes.

Laboratory evaluation of all Candidate genes in these areas requires exorbitant financial and time costs. Therefore, computational methods to prioritize candidate genes in these areas before laboratory methods can make a dramatic difference in detecting pathogenic genes.

Lack of properly integrated data and accuracy can be named as weaknesses of these methods. In this research, an attempt will be made to cover the weaknesses of the previous methods.

A study of what has been done so far shows that the development of a global network-based methodology combined with integrating data from different and heterogeneous data sources to create a gene ranking system with the possibility of automatically extracting new knowledge is essential.

7 Appendix A

```
for (int k = 1; k <= sad100; k++){
  tempTP = 0;
  tempFP = 0;
  tempTN = 0;
  tempFN = 0;
  if (k < seedindex){
    tempTP = 0;
    tempFP = k;
    tempTN = sad100 - k - 1;
    tempFN = 1;
    TP[k] = TP[k] + tempTP;
    FP[k] = FP[k] + tempFP;
    TN[k] = TN[k] + tempTN;
    FN[k] = FN[k] + tempFN;}
  Else {
    tempTP = 1;
    tempFP = k - 1;
    tempTN = sad100 - k;
    tempFN = 0;
    TP[k] = TP[k] + tempTP;
    FP[k] = FP[k] + tempFP;
    TN[k] = TN[k] + tempTN;
    FN[k] = FN[k] + tempFN;}
}//for k loop
}//for each file

for (int kk = 1; kk <= 100; kk++){
  Y[kk]=(double)TP[kk]/(double)(TP[kk]+FN[kk]);
  X[kk]=1-((double)TN[kk]/(double)(TN[kk]+FP[kk]));}
double AUC = 0.0;
for (int n = 1; n <= 100 - 1; n++) {
  AUC = AUC + (double)((Y[n] + Y[n + 1])
                            *(X[n + 1] - X[n])) / 2.0;}
```

8 References

[1] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.M. Agapow, M. Zietz, M.M. Hoffman, and W. Xie, "Opportunities and obstacles for deep learning in biology and medicine". *Journal of The Royal Society Interface, 15*(141), 2018. p.20170387.

[2] M.N. Weedon, L. Jackson, J.W. Harrison, K.S. Ruth, J. Tyrrell, A.T. Hattersley and C.F. Wright, "Use of SNP chips to detect rare pathogenic variants: retrospective", population based diagnostic evaluation. *bmj*, *372*. 2021.

[3] S. Erten and M. Koyutürk, "Role of centrality in network-based prioritization of disease genes". In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 8th European Conference, EvoBIO 2010, Istanbul, Turkey, April 7-9, 2010. Proceedings 8* (pp. 13-25). Springer Berlin Heidelberg. 2010.

[4] A.W. Kurian, K.C. Ward, A.S. Hamilton, D.M. Deapen, P. Abrahamse, I. Bondarenko, Y. Li, S.T. Hawley, M. Morrow, R. Jagsi, and S.J. Katz, "Uptake, results, and outcomes of germline multiple-gene sequencing after diagnosis of breast cancer". *JAMA oncology*, *4*(8), 2018. pp.1066-1072.

[5] Mu, W., Li, B., Wu, S., Chen, J., Sain, D., Xu, D., Black, M.H., Karam, R., Gillespie, K., Hagman, K.D.F. and Guidugli, L., 2019. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. *Genetics in Medicine, 21*(7), pp.1603-1610.

[6] Nicolosi, P., Ledet, E., Yang, S., Michalski, S., Freschi, B., O'Leary, E., Esplin, E.D., Nussbaum, R.L. and Sartor, O., 2019. Prevalence of germline variants in prostate cancer and implications for current genetic testing guidelines. *JAMA oncology*, *5*(4), pp.523-528.

[7] Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W. and Liu, B., 2018. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases*, *5*(2), pp.77-106.

[8] Pan, X., Hu, X., Zhang, Y.H., Chen, L., Zhu, L., Wan, S., Huang, T. and Cai, Y.D., 2019. Identification of the copy number variant biomarkers for breast cancer subtypes. *Molecular Genetics and Genomics*, *294*, pp.95-110.

[9] Yadav, S. and Couch, F.J., 2019. Germline genetic testing for breast cancer risk: the past, present, and future. *American Society of Clinical Oncology Educational Book*, *39*, pp.61-74.

[10] Kurian, A.W., Ward, K.C., Abrahamse, P., Bondarenko, I., Hamilton, A.S., Deapen, D., Morrow, M., Berek, J.S., Hofer, T.P. and Katz, S.J., 2021. Time trends in receipt of germline genetic testing and results for women diagnosed with breast cancer or ovarian cancer, 2012-2019. *Journal of Clinical Oncology*, *39*(15), p.1631.

[11] Chen, L., Zeng, T., Pan, X., Zhang, Y.H., Huang, T. and Cai, Y.D., 2019. Identifying methylation pattern and genes associated with breast cancer subtypes. *International journal of molecular sciences*, *20*(17), p.4269.

[12] Landrith, T., Li, B., Cass, A.A., Conner, B.R., LaDuca, H., McKenna, D.B., Maxwell, K.N., Domchek, S., Morman, N.A., Heinlen, C. and Wham, D., 2020. Splicing profile by capture RNA-seq identifies pathogenic germline variants in tumor suppressor genes. *NPJ precision oncology*, *4*(1), p.4.

[13] Yadav, S., Hu, C., Hart, S.N., Boddicker, N., Polley, E.C., Na, J., Gnanaolivu, R., Lee, K.Y., Lindstrom, T., Armasu, S. and Fitz-Gibbon, P., 2020. Evaluation of germline genetic testing criteria in a hospital-based series of women with breast cancer. *Journal of Clinical Oncology*, *38*(13), p.1409.

[14] Federici, G. and Soddu, S., 2020. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *Journal of Experimental & Clinical Cancer Research*, *39*, pp.1-12.

[15] Weitzel, J.N., Neuhausen, S.L., Adamson, A., Tao, S., Ricker, C., Maoz, A., Rosenblatt, M., Nehoray, B., Sand, S., Steele, L. and Unzeitig, G., 2019. Pathogenic and likely pathogenic variants in PALB2, CHEK2, and other known breast cancer susceptibility genes among 1054 BRCA-negative Hispanics with breast cancer. *Cancer*, *125*(16), pp.2829-2836.

[16] Fine, R.S., Pers, T.H., Amariuta, T., Raychaudhuri, S. and Hirschhorn, J.N., 2019. Benchmarker: an unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. *The American Journal of Human Genetics*, *104*(6), pp.1025-1039.

[17] Chen, J., Bardes, E.E., Aronow, B.J. and Jegga, A.G., 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, *37*(suppl_2), pp.W305-W311.

[18] Seelow, D., Schwarz, J.M. and Schuelke, M., 2008. GeneDistiller—distilling candidate genes from linkage intervals. *PloS one*, *3*(12), p.e3874.

[19] Nitsch, D., Tranchevent, L.C., Goncalves, J.P., Vogt, J.K., Madeira, S.C. and Moreau, Y., 2011. PINTA: a web server for network-based gene prioritization from expression data. *Nucleic acids research*, *39*(suppl_2), pp.W334-W338.

[20] Tranchevent, L.C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D. and Moreau, Y., 2016. Candidate gene prioritization with Endeavour. *Nucleic acids research*, *44*(W1), pp.W117-W121.

[21] Trifu, S.C., Vlăduţi, A. and Trifu, A.I., 2020. Genetic aspects in schizophrenia. Receptoral theories. Metabolic theories. *Romanian journal of morphology and embryology*, *61*(1), p.25.

[22] Azadifar, S. and Ahmadi, A., 2022. A novel candidate disease gene prioritization method using deep graph convolutional networks and semi-supervised learning. *BMC bioinformatics*, *23*(1), p.422.

[23] Hutz, J.E., Kraja, A.T., McLeod, H.L. and Province, M.A., 2008. CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *32*(8), pp.779-790.

[24] Danis, D., Jacobsen, J.O., Carmody, L.C., Gargano, M.A., McMurry, J.A., Hegde, A., Haendel, M.A., Valentini, G., Smedley, D. and Robinson, P.N., 2021. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *The American Journal of Human Genetics*, *108*(9), pp.1564-1577.

Article submitted 11 July 2023. Accepted at 19 August 2023. Published at 30 September 2023.