ORIGINAL RESEARCH PAPER

Medical Science

KEY WORDS: Breast Cancer, Gene Expression, Blood-Based Diagnostic Biomarker, Gene Expression Omnibus (GEO), STRING, Networks Analysis, Pathway

ANALYSIS OF GENE EXPRESSION DATASETS TO DISCOVER BLOOD-BASED BIOMARKERS IN BREAST CANCER

Vaishnavi Manivannan

ABA Oman International School, Muscat, Sultanate Of Oman

ABSTRACT

Breast cancer has an immensely hazardous impact on the world population. Despite advances in surgery, radiation and chemotherapy have been prevalent in the recent past, there still exists a need to study new biomarker (driving) genes to contribute to the development of personalized cancer treatment and drugs. In this study, we aim to analyze gene expression datasets for common differentially expressed genes (cDEGs) in the blood of stage 0-1 Breast cancer patients. Datasets were collected from the public Gene Expression Omnibus (GEO) repository. Upon analysis, 23 DEGs passed the cut-off criteria (p-value of < 0.5 and log fold change value of > 1.25). Common genes were identified from at least two out of the three datasets. In order to identify network, pathway characteristics and hub genes, computational tools of STRING and Jvenn were applied to a protein interaction network. Upon careful analysis and literature review, DDX6 (DEAD-Box Helicase 6) was found as a potential novel biomarker and warrants further study. Literature review confirmed this gene had been identified in relation to other forms of cancer (excluding breast cancer) in previous studies, thus showing novelty in relation to Breast cancer. Studying these 23 genes could illuminate a new direction of the development of effective breast cancer treatment. Overall, in this study we present findings of different insights on molecular mechanisms of Breast cancer and provide greater confidence on which genes are differentially expressed in Breast cancer.

INTRODUCTION

Breast cancer (BC) is the most common type of cancer found within women and has a high mortality rate. It is occasionally found in men, but affects "12% of women worldwide, and 30% of them die from it" (NCBI, 2018). With such a high death rate, it is essential to develop better systems for early detection. As with many types of cancer, breast cancer is tumorous and grows to affect multiple parts of the body, making it difficult to develop effective treatments that completely eradicate the cancer cells from a body. From previous studies, cancer driver gene biomarkers have been found from microarray data and experimentation. However, these approaches are known to be flawed because of a poor quantity of samples, which has created inconsistent and inaccurate cancer biomarkers. This means that the genes identified in one study may not be as significant in other studies. This calls for an alternate approach of analyzing gene expression data in order to identify biomarkers. This is where computational analysis of gene expression datasets comes in. Gene expression datasets can then be analyzed in different ways in order to find common cancer driving genes. These include; differential expression analysis, network analysis, gene ontology analysis, pathway analysis, as well as comparing expression levels of one or more genes from different samples (which will be discussed predominantly in this paper). In the present study, microarray data from three separate studies by different authors were analyzed to identify biomarkers for early stage breast cancer. The studies all contained gene expression data based on the blood of patients who either had breast cancer or were normal controls. Differentially expressed genes identified across studies were considered to be of special interest and their interactions were studied further.

Existing Biomarkers

There are 4 major types of biomarkers used in the cancer diagnostic and treatment pipeline. These are 1) Diagnostic, 2) Prognostic, 3) Predictive and 4) Pharmacodynamic. This study will specifically focus on identifying diagnostic biomarkers for breast cancer. From current scientific studies, we know that "germline mutations in TP53 and PIK3CA are the most common driver genes of breast cancer" (BioMed Central, 2021). This is because the TP53 gene is a crucial 'tumor suppressor gene which also rectifies DNA damage' (BioMed Central, 2021) . On the other hand, PIK3CA is crucial in controlling cell division and replication. Through a technique called gene expression profiling (monitoring and analyzing the activity levels of different genes driving a particular cancer, breast cancer in this study's case), different subgroups of breast cancer has been classified: "SMAD4, ERBB2, KRAS, ARID1A, CDKN2A, PBRM1, KDM6A, MEN1, FOXP1, USP9X, BAP1" (Cancer Center, 2019). These gene biomarkers help guide doctors in coming up with a specialized/targeted treatment plan. While there are other biomarkers that have been identified, they are yet to undergo laboratory testing to confirm its validity and reliability of being used as a biomarker. Despite advances made, the aforementioned current widely used methods is not cost or time effective.

METHODOLOGY

Data Selection

The study will use secondary data analysis. These include the Gene Expression Omnibus (GEO) and Array Express. The Gene Expression Omnibus Dataset (GEO) is an online National Center for Biotechnology Information (NCBI) database containing gene expression datasets from conducted studies. The keywords "breast cancer" were searched for with the specifications of "expression profiling by array" and "Homo sapiens." By doing so, independent datasets would be chosen given it is blood-based and focused on the species of homo-sapiens. Upon searching the aforementioned keywords, the selection criteria GSE27567, GSE65517 and GSE27473 were selected for this study. The number of cancer and control samples in the 3 chosen datasets.

Table 1: Gene Expression Omnibus Dataset Information On The Number Of Cancer And Control Samples Analyzed

Dataset	Number of cancer samples	Number of control samples
GSE27567	94	31
GSE65517	4	3
GSE27473	3	3
Total	101	37

Differential Expression Analysis

GEO2R is a browser-based software that processes gene expression values and outputs a table of differentially expressed genes (DEGs) between two user-defined groups (breast cancer and control, in this case). In total, several thousand genes were deemed statistically significant by GEO2R for each dataset. T tests were used to determine P values. GEO2R was also used to calculate fold changes for

each gene in the context of breast cancer expression values vs. control values. The fold change is a ratio of the average expression value of a gene in one group divided by the average expression value in a different group. Fold change can be greater or less than one; some genes are overexpressed while others are underexpressed in breast cancer compared to normal controls. GEO2R was also used to verify a normal distribution of gene expression values. No outliers were present. After the lists of DEGs were obtained they were placed in Google Sheets and processed.

DEGs with p values of greater than 0.05 were removed as these values accept the null hypothesis. A small P value (< 0.05) indicates evidence of differential expression, and a large P value indicates a false positive. This is why, I deleted the rows recorded a P value of over 0.05 as it is insignificant. The log fold change value indicates how overexpressed or underexpressed a gene is in the experimental and control category. Positive logFC values mean the gene is overexpressed while a negative logFC value means the gene is underexpressed. A value of exactly 1 means no difference. With this relationship, the remaining DEGs were sorted into two categories: overexpression and underexpression. This was done by sorting the rows in an increasing order of the logFC column. Next, only genes with fold change of greater than 1.25 were kept because this is the cut-off value. After applying the cut-off criteria (p value <0.05 and fold change > 1.25) for all datasets, data was processed.

Identification of common genes: The online tool Jvenn was used to create a Venn diagram of the genes contained in each of the three studies. Genes present in at least two out of three studies were saved while genes found in only one were removed from the list.

Network analysis with STRING and PANTHER: The software of STRING was used to analyze the final list of DEGs by visually displaying the input genes and their interactions. This tool was also used to identify significant Gene Ontology processes and pathways. The PANTHER tool was used to detect enriched biological processes and molecular functions between the common genes. STRING was also used in the process of text-mining, which involves analyzing the most recent and prominent literature data that comprises of the genes discovered in this study. The text mining method through scientific literature is simple and feasible. The specific gene terms and breast cancer will be keywords when searching through the literature database, then will be consulted with the studies found to see if there is a strong correlation between the identified genes and aim of the respective studies.

RESULTS

Figure 1: Venn diagram that shows the intersecting genes between the three datasets of GSE27567, GSE27473 and GSE65517.

Analysis with JVENN tool: In GSE27567, 3 genes passed the given criteria in GSE27473, 1047 genes and in GSE65517 473 genes passed the set criteria. There was a considerable amount of overlap between the three datasets. 22 genes (ASPM, TMEM45A, IFIT3, RPH3AL, IL7, ST3GAL3, DYSF, NAV2, AMPD3, HLX, FMNL2, ETV4, ZNF765, CASP4, PRKCDBP, MBD1, ETS1, RPS6KA2, CDK14, DDX60L, FHL2, F2RL1) were found in two out of the three datasets and 1 was found in all three. This gene was the DDX6 gene. As mentioned in the 2. Methodology section, 2.3 Identification of common genes section, the software of Jvenn was used to find commonalities—this is shown in Figure 1.

The common gene out of all three datasets was the DDX6 (DEAD-Box Helicase 6) gene. With current research, this gene is known to be associated with non-cancerous conditions centering Intellectual Development Disorders. "Among its related pathways are Deadenylation-dependent mRNA decay and Processing of Capped Intron-Containing Pre-mRNA. Gene Ontology (GO) annotations related to this gene include nucleic acid binding and protein domain specific binding" (Gene Cards, n. d). | After using JVenn for finding the common genes, String DB software was used to generate a network model of predicted associations for a particular group of proteins that connect to a specific gene (DDX6). In the network, the thickness of the lines between different proteins indicate the degree of confidence prediction of the interaction.

As shown in Figure 2 below, denote the STRING network representing the protein interactions of the 22 genes and also the protein interactions of the gene DDX6, respectively. The color of the lines connecting the different genes signify the strength of the interaction between them. A pink line, for example, indicates the gene's experimental demonstration. Isolated genes are removed from the diagram for visual clarity. also the protein interactions of the gene DDX6, respectively. The color of the lines connecting the different genes signify the strength of the interaction between them. A pink line, for example, indicates the gene's experimental demonstration. Isolated genes are removed from the diagram for visual clarity. The minimum confidence score for whether a protein interaction existed was set to 0.400 under the Settings. There were 42 connections between proteins instead of an expected 105, indicating that the network has significantly more interactions than predicted of a random sample (p value = 0.004).

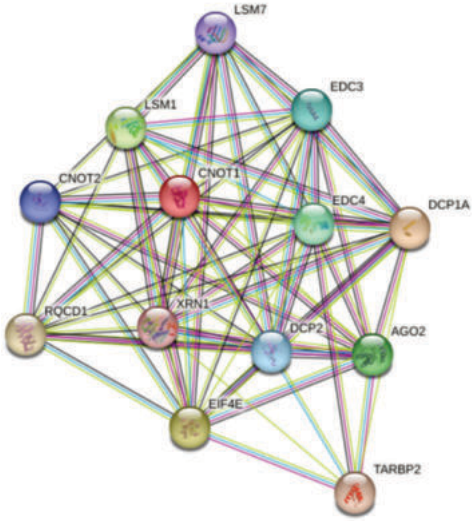

Figure 2: String Network Of Protein Interactions Between The 13 Genes Of The Ddx6 (dead-box Helicase 6) Gene. Lsm1 And Cnot1 Gene Had The Highest Number Of Interactions With The Other Proteins. Isolated Genes Were Removed From The Diagram For Clarity.

CONCLUSION

Analysis of three microarray datasets on breast cancer gene expression yielded 23 differentially expressed genes that are found across studies. These genes passed the criteria for p-value and fold change (set criteria for valid genes) and are attractive for further research as they are biologically relevant. While 23 genes were present in two of three studies, only one—DDX6 (DEAD-Box Helicase 6) was found as a commonality between all three datasets. As seen in Figure 1, through the String DB pathway analysis—LSM1, DCP2, DCP1A, EIF4E, AGO2, CNOT1, EDC3, EDC4, PATL1 and LSM14A all had at least eight interactions with the gene of DDX6, indicating that these protein interactions may play an important role in Breast cancer. LSM14A was located in the center of a cluster of other genes, posing more of a significance than rest of the connected genes. The tool of STRING was used to "text-mine" (shown in Appendix 4) through research papers published from 2000 till date in order to check the novelty of this study and whether or not DDX6 had already been classified as a breast cancer biomarker. This text mining step is also beneficial in evaluating whether the gene of DDX6 has been investigated in general cancer studies. Figure 3 shows a mention of either the gene DDX6 or its protein/gene pathways within a cancer study. These papers don't centrally focus on DDX6, instead mention it alongside a central idea. This indicates more validity and reliability in conducting further experimental studies to analyze the genetic composure of DDX6, in order to confirm it's role as a biomarker in breast cancer.

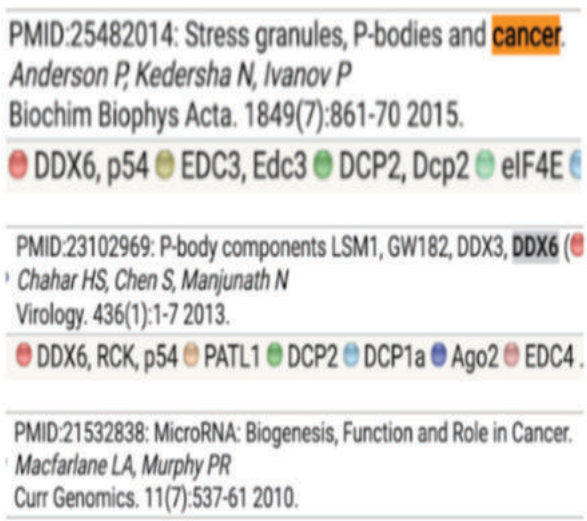

Figure 3: STRING data mining through scientific literature to detect cancer from the presence of DDX6.

EVALUATION

Strengths of this study include the rigor applied to differentially expressed gene selection. To qualify for the final list, a gene had to have a fold change of at least 1.25, a p value of < 0.05, and be present in datasets. This rigid criteria shows how the DEGs selected from the analysis is of high validity and reliability. Secondly, another strength is the sample size of the datasets. Due to its varied intensity levels, Breast cancer can vary considerably between patients. In addition, it is essential to keep the stage of the breast cancer of the patient controlled throughout the selected datasets. These restrictions are followed when selecting the datasets from the Gene Expression Omnibus GEO platform. At late stages, there are often different gene expression differences that may not necessarily be present at early stages. With the great sample size used in this study, there is greater assurance of obtaining a general view of the gene expression of breast cancer than with just a few samples, for example, in the case of GSE27473.

Limitations of this study are that no normalization was performed between datasets. Normalization refers to when the scaling for a specific column in a dataset is altered without changing the range of the values. This aims to makes the data more easy to navigate (which reducing congestion) and also to get the data points to be on a similar scale.

However, this is not common with cross-study gene expression analyses, hence it might not have posed as a reasonable problem. Another limitation is that there was no distinction between male and female samples in the analysis within the datasets on Gene Expression Omnibus. There may be significant differences in gene expression and breast cancer pathology between sexes that were ignored in this study, which could cause serious concerns in the real-world context. Biomarkers that may work well for females may not be effective for males. Another important aspect to note is that the research only compared healthy normal controls and early stage breast cancer patients, meaning there are other factors in relation to time not taken into account. Thus, it is not known whether the DEGs found are specific to breast cancer, overall (while considering all the different stages).

Future experimental studies are needed to validate these markers in bigger datasets, to determine their role in breast tumorigenesis, develop liquid biopsy/biosensor based approaches, and move this information to clinic for early identification of breast cancer risk. In addition, further experimental and molecular studies are required especially in cell lines and animal models to show conclusively whether or not each or a combination of these markers can be utilized as indicators of breast cancer risk without having observable effects on breast cancer cells or can have other roles at the earlier stages of carcinogenesis. Overall, our analysis offers novel biomarkers for further validation and functional characterization.

REFERENCES

1. Biological interpretation of gene expression data. (n.d.). EBI. Retrieved October 2, 2022, from https://www.google.com/url?q=https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/biological-interpretation-of-gene-expression-data-2/&sa=D&source=docs&ust=1665880628536433&usg=AOvVaw1gkt4Nn4_SZcYKqsB3Oemr
2. biomarker. (n.d.). National Cancer Institute. Retrieved October 16, 2022, from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker
3. Gene Expression. (n.d.). Nature. Retrieved March 16, 2022, from https://www.nature.com/scitable/topicpage/gene-expression-14121669/
4. Is fold change of value 1.5 (log2FC = 0.58) a significant value ? (n.d.).
5. BioStars - Bioinformatics Explained. Retrieved May 15, 2022, from https://www.google.com/url?q=https://www.biostars.org/p/373655/ &sa=D&source=docs&ust=1665880701333584&usg=AOvVaw3Xd0cj2-beagku9gYAN0R_