INTRODUCTION

Breast cancer is the most commonly diagnosed malignancy among females. While decrease in both, breast cancer incidence and mortality, have been apparent in recent years, the societal and economic impact of this malignancy continues to be enormous [1]. The cases of incidence were 1.8 million in 2013 and 464 thousand deaths approximately [2]. Nearly, 30% of all cancers in women occur in breast both in the developed and the developing world [3]. The genetic abnormalities such as variations in high-penetrance genes play a major role in about 90% of breast cancer cases. Several risk factors for breast cancer have been identified. Some mutations particularly in BRCA1, BRCA2, p53, PTEN, ATM, NBS1, and LKB1 result in a very high risk for breast cancer [4].

Breast cancers are of two different types, invasive and non-invasive. Invasive cancers spread to other tissues in the breast from the milk ducts, whereas the non-invasive cancers do not invade other tissues in the breast. The non-invasive breast tumors are referred to as “in situ.” These are classified as ductal carcinoma in situ (DCIS) or intraductal carcinoma and lobular carcinoma in situ [4].

DCIS is characterized by malignant epithelial cells confined to the ductal system of the breast, without evidence of invasion through the basement membrane into the surrounding stroma [5]. Once thought to be a rare breast lesion, DCIS now constitutes 20% of all newly diagnosed breast cancer cases [http://seer.cancer.gov, Accessed October 2013; http://www.cancer.org, Accessed June 1, 2012]. Invasive breast cancer (IBC) starts in a milk duct of the breast, breaks through the wall of the duct and grows into the fatty tissue of the breast. At this point, it may be able to spread (metastasize) to other parts of the body through the lymphatic system and bloodstream [5]. Invasive breast carcinoma constitutes 70-85% of the incidence; the remaining 15-30% are in situ carcinomas, 80% of which are DCIS [http://www.cancer.org, Accessed June 1, 2012].

The factors that stimulate the breast cancer risk include gender, age, family history and additionally alcohol intake, dietary fat, obesity in postmenopausal age, and hormonal stimulations. These factors are said to have increased the progression of breast cancer along with the individual factors almost half a century. The dramatic increase in breast cancer research and its prevention has shown positivist approach in the current years [6].

With the advent of microarray technology, the procedure to measure gene expression on a genome-wide scale has transformed cancer biology by providing the tools to measure differences in diseases [7]. This technology utilizes differential gene expression patterns in cancer cells and normal cells or those of other subtypes of cancer to identify the genes that are over-expressed and under-expressed [8]. However, the analysis produces a large amount of data, which is challenging to interpret. With the employment of modern computational and statistical analysis packages and bioinformatics tools, the data analysis has been greatly flexible in the recent years. The microarray technology has been applied to a range of applications, including discovering novel disease subtypes, developing new diagnostic tools, and identifying underlying mechanisms of disease or drug response [9].
In this work, we studied the gene expression profiling of breast cancer samples from Gene Expression Omnibus (GEO) database. The dataset GSE41194 was retrieved from GEO, to investigate differential gene expression in DCIS and IBC. Gene expression profiling of DCIS and IBC was performed to discover uniquely expressed genes that also regulate the progression. Our study also focused on identifying pathways associated with the genes, which enables to develop novel treatments for DCIS and IBC.

METHODS

Data quality check for the samples in the dataset
To check the quality and detect the outliers within the samples in the dataset, diagnostic plots such as boxplots and density plots were plotted. These plots give a quick view of the normalized log2 intensities.

Gene expression in DCIS and IBC
To investigate the differential expression in DCIS and IBC, the dataset GSE41194 deposited by Lee et al. [10], titled differentially expressed genes (DEGs) regulating the progression of DCIS to IBC (Group 1) was downloaded from GEO database [11]. The dataset contains 26 DCIS samples and 24 IBC samples. The platform used was GPL8300 [HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array. The original files (raw data) and the platform probe annotation files were downloaded.

Identification of DEGs
The original data were classified as DCIS and IBC groups and were analyzed using R software (v.3.0.1) [12] and Bioconductor (v.2.14) [13]. The multichip normalization method multiarray average was used for background correction, normalization across the chips, and summarization of probe level data [14]. Finally, Limma–Linear Models for Microarray Data [15], linear regression model software, were used to compare the differential expression on different classes of chips. To identify the differentially expressed genes in DCIS and IBC, an adjusted *p value ≤0.05 was used as the cut-off criterion. Furthermore, to filter the differentially expressed probe sets, a *fold change (FC)≥2 was used to identify upregulated genes and ≤−2 for downregulated genes.

Gene ontology (GO) of DEGs
To investigate the DEGs at a functional level: Primarily, Database for Annotation, Visualization and Integrated Discovery (DAVID)-v.6.7 [16] was used to functionally interpret gene lists, to analyze the GO classification of terms [17], for identification of cellular components (CC), biological process (BP), and molecular function (MF) and for visualizing genes and mapping pathways Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] was used. The DEGs for stronger gene enrichment analysis were chosen with an *EASE Score Threshold of ≤0.1 for the maximum probability and a default *Count threshold (minimum count) of 2 for including the minimum number of genes for the corresponding GO term.

RESULTS

Quality analysis of samples in the dataset
The quality analysis involves the assessment of the data and detection of the outliers. In this analysis, boxplots and histograms were plotted to see whether the samples in the dataset had any outliers. The boxplot of the raw data (Fig. 1) represents the distribution of log2 intensities across all the samples. The boxplot of normalized signal intensities across all samples provides a certainty that the normalization step was accomplished (Fig. 2).

The density plot shows the biased log2 intensity distribution for all the samples (Fig. 3). The histogram obtained after normalization makes the distributions essentially the same across all the samples (Fig. 4).

Identification of DEGs
The limma package was used to build model matrix with defined contrasts and an adjusted false discovery rate to analyze the gene
expression profiles of DCIS and IBC. The analysis identified 72 genes that were found to be differentially expressed with the adjusted p≤0.05 and |FC|≥2 and ≤−2. The |FC|≥2 and ≤−2 revealed 38 probe sets that were upregulated and 34 probe sets that were downregulated.

GO clustering and pathway enrichment of DEGs

The functional classification of the obtained 72 DEGs was performed with the online biological classification tool-DAVID. The gene list was submitted with Affymetrix Human U133 chip as background and was provided for enrichment calculation. An *Fisher exact p value ≤0.1, was used for strong gene-enrichment. The count threshold (minimum count) of 2 was used to retrieve minimum gene counts belonging to a GO term with its categories (classifications) - BP, CC, and MFs. The functional annotations of gene classifications, with their GO terms, p-value, count, and percentages that present study identified are detailed in Table 1. The DAVID analysis revealed six genes that were significantly associated with GO terms and pathways. The GO associated with the genes are shown in Table 2. The KEGG pathway associations for the obtained genes are reported in Table 3. Further investigation on these genes and pathways pave a novel way for developing new therapies for treating patients with breast cancer.

**DISCUSSION**

Breast cancer accounts the principle cause of death among women, with high incidence rates in Australia/New Zealand, North America, and several European countries [19,20]. It is estimated that 1.67 million breast cancer cases have been diagnosed in 2012, according to the Global Cancer Observatory series of the International Agency for Research on Cancer [21]. The steady increase in the morbidity of breast cancer in the recent years indicates a need for additional research on this disease.

Breast cancer has been described as an alarming health problem in India. It is the second most common cancer. A survey carried out by the Indian Council of Medical Research in the metropolitan cities, viz., Delhi, Mumbai, Bangalore, and Chennai; from 1982 to 2005, has shown that the incidences of breast cancer have doubled. Over the years, the incidences of breast cancer in India have steadily increased and as many as 100,000 new patients are being detected every year. A 12% increase has been registered by cancer registries from 1985 to 2001, which represented 57% rise of cancer burden in India [22].

The differential gene expression analysis of DCIS and IBC identified 72 genes with a significant p value ≤0.05; in which 38 probe sets were upregulated, and 34 probe sets were downregulated. These DEGs retrieved are important for investigating the mechanism of disease development from DCIS to IBC. It is well known that breast cancer treatment has severe side effects which enable to find better chemotherapeutic agents [23]. The analysis of differentially expressed genes and their associated annotations provide useful information with genes function in all GO categories. These may be useful for determining and understanding the specific gene expression in the development of disease targets for the treatment of DCIS and IBC.

The results of GO functional annotation and pathway enrichment analysis of DEGs retrieved 6 genes and their associated pathways that were significantly enriched with GO terms, classifications and were associated with Mitogen-Activated Protein Kinase (MAPK) signaling.
pathway, neurotrophin signaling pathway, cysteine and methionine metabolism, tyrosine metabolism, phenylalanine metabolism, tyrosine and tryptophan biosynthesis cellular signaling pathways, and others. The MAPK pathway identified in our analysis is one of the principal targets for treating breast cancer [24]. This pathway is involved in various cellular functions, including cell proliferation, differentiation, and migration signaling [25]. Hence, it is necessary to carry out additional research for identifying potential targets as therapeutic agents that may directly or indirectly be involved in treating breast cancer.

CONCLUSION

Microarrays emerged as large-scale experimental studies to generate expression of thousands of genes parallelly. This technology makes biological observations more significant from a statistical point of view. Our study focused on analyzing and understanding breast tumor types - DCIS and IBC, as these data can provide us a wealth of information on the genetic susceptibility of disease through which decisive steps can be taken to translate these findings to clinical care. Gene expression profiling of breast tumors enables us to have a better understanding of tumor type and what markers certain tumors may have. Identifying gene profiles for DCIS and IBC tumors allows us to better group and classify these tumors. This enables development of better drug and treatment procedures. The gene expression affected by complex biochemical pathways and signaling events can be studied eventually.

ACKNOWLEDGMENTS

We would like to thank SRM University for all the support extended toward this research.

REFERENCES