The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/DG6TK

TRANSCRIPTOMICS ANALYSIS OF BREAST CANCER TISSUES: AN IN-SILICO APPROACH USING MACHINE LEARNING FEATURE SELECTION ALGORITHMS

ALPNA SHARMA[1], NISHEETH JOSHI [2] and VINAY KUMAR [3]

[1, 2]Department of Computer Science, Apaji Institute, Banasthali University, India.
[3]Ex Scientist GOI, Ex Professor, VIPS – Vivekananda Institute of Professional Studies.
Email: alpna@vips.edu[1], jnisheeth@banasthali.in[2], vinay5861@gmail.com[3]

Abstract

The most frequent cancer in women and the second most common cancer overall among newly diagnosed cases is breast cancer. Local invasion and metastasis are factors that precede the majority of cancer fatalities, with metastasis accounting for 90% of deaths, but very little is known about the molecular causes of invasion and metastasis. Thus exposing the underlying causes of this condition at the Transcriptomics level can lead to a novel treatment approach for Breast Cancer. To identify underlying differences between epithelial breast cancer tissues (TEC), stromal breast cancer tissues (SCC), normal control epithelial breast cancer tissue samples (EN), and normal control stromal breast cancer tissue samples (SN) at the Transcriptomics level, the total RNA microarray processed data from GEO for breast cancer patients was analyzed. The transcriptional profiles of 64 samples, including 28 TEC, 28 SCC, 5 EN, and 5 SN controls received from the NCBI-Bio project, were therefore subjected to various bioinformatics analysis in the current work (PRJNA107497). First, exploratory data analysis based on gene expression data using principal component analysis (PCA) depicted distinct patterns between TEC vs EN and SCC vs SN samples. Subsequently, the Welch's T-test differential gene expression analysis identified 22277 significantly differentially expressed genes (Fold change (>= 1.5), p.adj <value 0.1) between these conditions. This study reveals the genes like COL11A1, COL1A1, COL1A2, COL3A1, COL5A1 and COL5A2 as the key features that may substantially contribute to metastasis of breast cancer from epithelial cells to stromal cells in the mammary glands. As a result of the up-regulated and down-regulated genes, this study was also able to pinpoint the affected biological pathways for both the SCC vs. SN samples and the TEC vs. TN samples. This most definitely offers an important clue regarding the root of the fatal metastatic cancer problem. Ultimately, the findings provided here offer fresh perspectives on breast cancer metastasis.

Keywords: Breast cancer, Machine Learning, Differential gene expression, KEGG pathway analysis, PCA, Heat maps, Dendrogram

1. INTRODUCTION

Due to its high mortality and morbidity rates, breast cancer is one of the main health issues for women (1). Even with adjuvant chemotherapy, the five-year survival rate for metastatic breast cancer is less than 30%. (2). Breast cancer (BC) is the most common malignancy that affects women worldwide. In 2020, it will surpass lung cancer as the most prevalent type of cancer globally, with a projected 2.3 million new cases annually, or 11.7% of all cancer cases (3). Epidemiological studies predict that there will be more than 2 million cases of BC worldwide by the year 2030. (4). between 1965 and 1985, the incidence in India increased significantly—nearly by 50%. In India, there were an estimated 118000 incident cases in 2016 (95% confidence interval: 107000–130000), 98.1% of whom were female, and 526000 prevalent cases (474000 to 574000) (3, 4). Every state in the nation has seen an increase in the age-