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-

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

standardized incidence rate of BC in females over the past 26 years, which is up 39.1% (95% confidence interval, 5.1 to 85.5) (4).

Local invasion and metastasis are factors that precede the majority of cancer fatalities, with metastasis accounting for 90% of deaths from solid tumors (5). Unfortunately, little is known about the molecular causes of invasion and metastasis (5). Malignant epithelial cells must infiltrate into the surrounding breast stroma through the basement membrane extracellular matrix (ECM) in order for DCIS to proceed to stage I breast cancer. Cancer cells gain the ability to infect nearby vascular structures and spread once invasion has taken place and they have entered the stroma (6). Studies in genetics and cell biology have demonstrated that the tumour stroma is necessary for tumour growth and progression in addition to the altered epithelial cells (7, 8). Fibroblasts, adipocytes, the ECM, and blood and lymph arteries make up stromal tissue, all of which have been found to affect tumour development. As cancer spreads, the stroma within the tumour microenvironment is altered. These changes include fibroblast activation, ECM remodeling, and angiogenesis (9). These modifications are thought to be crucial in converting the stroma into a metastasis-supportive milieu.

The transcriptome of each of these compartments either at epithelial or stromal level must be studied individually in order to identify effects of each of these compartments on the cell and molecular level. Hence, there are studies that are concentrated at identifying an independent molecular signature of stromal tissue linked with cancer are constrained. According to current research employing laser capture microdissection (LCM), the metastasizing primary breast tumour is detected by the up-regulation of stroma specific genes together with the state of inactivation of tumor-epithelial specific genes and signals (9). Moreover, research and investigations were carried out utilising normal epithelium and stroma samples taken from patients undergoing reduction mammoplasty or surgical treatment for breast cancer, which were then dissected using the LCM method. These studies have additionally shown that the stromal microenvironment modifications are not present prior to the stage before carcinogenesis is initiated and are instead directly associated to cancer progression and metastasis (10). ED-A splice of fibronectin along with Alpha smooth muscle actin (SMA) have been identified to increase fibroblast activation in fibrosis and wound healing recently. Now, it has been discovered that fibroblasts are active in cancer, which is consistent with the idea that tumours are comparable to a persistent wound that doesn't heal (11). These activated fibroblasts, also known as cancer-associated fibroblasts (CAFs), resemble activated fibroblasts that are present in wounds and inflammatory regions in many ways (12). Because of their varied cellular origins and expression markers, CAFs are currently not precisely defined. These studies provide support for the notion that understanding the development of metastatic disease depends on understanding the transcriptome of the tumour microenvironment. Cancer associated fibroblasts (CAFs), which are different from regular fibroblasts and play a role in mediating tumour invasion and metastasis, are hypothesized to play a role in breast cancer, which is an epithelial cell phenomena (12, 13).Epithelial tumour cells can invade and spread throughout the body as a result of changed gene expression in the tumour stroma brought on by the progression of breast cancer. Furthermore, fibroblasts and epithelial cells cooperate to facilitate tumour invasion and metastasis. Owing to the about explanations, we therefore

ISSN 1533-9211

DOI 10.17605/OSF.IO/DG6TK

hypothesize that invasion by fibroblasts is caused partly by alterations in the ECM, which altogether creates a complicated framework for the angiogenesis and the migration of tumour epithelial cells (14).

Despite the fact that cancer-specific mortality has decreased as a result of therapeutic procedures such surgery, chemotherapy, radiation, endocrine therapy, and targeted therapy, there are still numerous therapeutic failures that lead to cancer recurrence, metastasis, and death (15). Understanding the molecular mechanisms by which reactive stromal fibroblasts affect cancer cells will help to improve therapeutic outcomes in the treatment of breast cancer.

Our transcriptomics study's main objective was to uncover the underlying causes of the discrepancies in the gene expression profiles of cancerous stromal (SCC) and cancer epithelial (TEC) tissues as compared to control samples. Our central concern was to identify the affected pathways at the Molecular, Biological and Cellular levels due to the differences in gene expression patterns seen among TEC and SCC patient tissue samples. Our main goal was to establish the function stroma plays in breast cancer invasion by examining the identified transcriptome differences between stromal and epithelial cells in normal breast tissue and breast cancer tissue samples. We looked for transcriptional errors to gain a better understanding of the growth, development, and progression of tumours in breast cancer. Future treatments may greatly benefit from identifying the underlying genes that cause breast cancer because the expression of TEC and SCC is tissue-specific and depends on the phenotypes in which it occurs.

Therefore, in the current study, we looked at the transcriptomics profiles of breast tissue samples taken during surgery from patients having invasive breast cancer surgically removed (n=56) [Cancer Epithelial Tissues=28; Cancer Stromal Tissues=28] and normal breast tissue samples taken from patients (n=10) to understand the affected Molecular, Biological, and Cellular pathways and their corresponding differentially expressed genes at the transcriptomics level showing We separated the impacted pathways as the disease progressed to later phases of breast cancer metastasis after identifying the relevant genes between these two experimental groups.

2. MATERIALS AND METHODS

2.1 Data sets

In this particular research study, the transcriptome data was extracted from the NCBI GEO server and the data set number was [GSE10797]. The dataset of this project was generated by Casey T, Bond J, Tighe S, Hunter T et al. and is published as a bio project on NCBI with the bio project accession number PRJNA107497. Transcriptome data contains microarray processed quantile normalized values using the Bioconductor package Lumi (version 2.32.0) of total RNA extracted from the stromal and epithelial breast cancer cells from the [HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array platform. The data set of the mentioned Bio project is depicted in Table 1 below.

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

Table 1: Datasets used in the present study

Disease State Sample Group Name	Number of Samples
Cancer Epithelial Tissues [TEC]	28
Cancer Stromal Tissues [SCC]	28
Normal Stromal Tissue [SN]	5
Normal Epithelial Tissue[EN]	5

2.2 Data Pre-Processing

Quantile-normalized signal data was used to process the microarray data. There was Gene IDs in the processed data. Using the SOFT family files, we mapped Gene IDs to gene symbols.

2.2.1 Exploratory Analysis

The exploratory study showed the comparison of the three groups of samples, comprising samples from healthy and diseased patients (Cancer Epithelial and Cancerous Stromal tissues). The following exploratory data analysis was carried out using the principal component analysis tool built within the Metaboanalyst Bioinformatics Server (https://www.metaboanalyst.ca/), in order to better understand the patterns in the data. The diversity between the data is discretely represented and visualized using PCA, a dimensionality reduction approach (15). PCA was performed in following three independent conditions: (1) All samples (Cancer Epithelial tissues vs Cancer Stromal tissues vs control samples); (2) Cancer Epithelial tissue vs Normal Epithelial Tissue; (3) Cancer Stromal tissues vs Normal Stromal Tissue; Followed by that, the PCA scatter plots were plotted to determine the patterns.

2.2.2 Differential Gene expression Analysis

Comparing the TEC and SCC samples with the Control samples, EN and SN, respectively, allowed for the differential Gene Expression (DGE) analysis to be carried out. The samples were subjected to a differential gene expression (DGE) analysis utilizing the Welch's T-test on the Metaboanalyst Software and the GEO2R built-in DEG function tool by NCBI GEO platform. To identify the relevant genes and offer statistical significance for variances with unequal variances, the Welch's T-test is a statistical analysis modification (15). When two groups have unequal sample sizes and variances, a Welch's test might be utilized (15). The threshold of (p.adj value<0.1, Fold change (>= 1.5)) was used to identify the potentially differentiating important genes (16). Apart from this, we used the feature selection algorithm on metaboanalyst to scrutinize the top 25 differentially expressed genes, so that one can get a better idea of the genes involved in the metastasis of Breast cancer.

2.2.3 Assessment of Discriminatory potential of significant genes

Then, using the chosen set of differentially expressed significant genes alone, statistical analysis such PCA, dendrograms, and H-Clustering was carried out in order to evaluate and illustrate the potential of the found significant genes in differentiating both classes of data. In order to comprehend the potential of significant genes frequent in differentiating the TEC, SCC, and Control samples based on their gene expression, H-clustering (Distance: Euclidean, Linkage: average) was carried out. The gene expression patterns among the various classes of

ISSN 1533-9211

DOI 10.17605/OSF.IO/DG6TK

samples were eventually depicted using heatmaps. Techniques like Heat-map and H-clustering make it easier to evaluate improved feature selection for the cutting-edge machine learning algorithms that have been proposed. Heatmap makes it simple to determine which features are most closely associated to the target variable, whereas H-clustering enables us to see how our data set clusters in order to see how features are chosen. Also, this suggests that there is a factor at the transcriptomics level that separates the genes into distinct heat map expression patterns and Dendrogram clusters.

2.2.4 Gene Enrichment Analysis

The Enrichr: Pathway analysis software's annotation module was used to do a gene enrichment analysis for Gene Ontology (GO) concepts in order to clarify the biological importance of the key genes. Also, the Enrichr software was used to find significantly expressed genes (p.adj value 0.05, Fold change (>= 1.5) that enriched the KEGG pathways (16,25). Enrichr is a web-based enrichment analysis application that is simple to use and intuitive and offers many sorts of visualization summaries of the collective functions of gene lists (16,25). Additionally, utilizing the Enrichr software-based platform, enriched pathways were also found and examined, notably for the significantly differentially expressed down-regulated and up-regulated genes.

3. RESULTS

To uncover the underlying gene signatures and biological pathways at the transcriptomics level, we therefore explored and analysed the transcriptomics data of cancer epithelial tissue, cancer stromal tissue, and healthy control epithelial and stromal samples using various bioinformatics techniques in the current proposed study. Figure 1 depicts the entire study's workflow in its entirety.

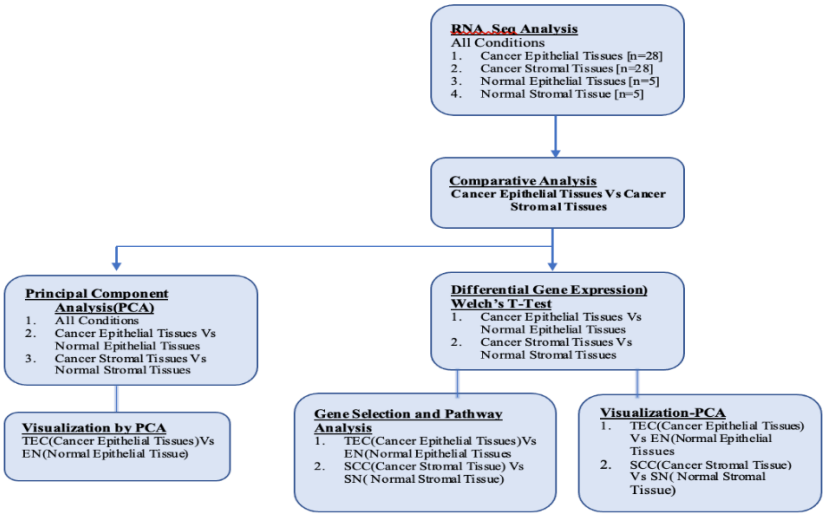

Figure 1: Work Flow of the study representing the key steps

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

3.1 Exploratory Data Analysis

Principal component analysis was used to retrieve the transcriptomics data (PCA). We used PCA to analyze the underlying variation across the three groups, which included samples from healthy control epithelial and stromal samples and cancer epithelial and stromal tissue. The fluctuation between all the circumstances, including TEC, EN, SCC, and SN, is shown in Figure 2A. The PC1 is 26%, the PC2 is 22%, and the PC3 is 5%. The PCA of this exploratory analysis makes it clear that something at the transcriptomics levels is responsible for the gene-level differences between the TEC, EN, SCC, and SN samples. Importantly, there are significant differences between the TEC and EN samples, with PC1 being 18.9%, PC2 being 11.4%, and PC3 being 9.4%, as seen in Figure 2B.Figure 2C represents variation between SCC and SN samples. The variation between 3 PCAs is given as PC1=20.3%, PC2=7% and PC3=4.8%.

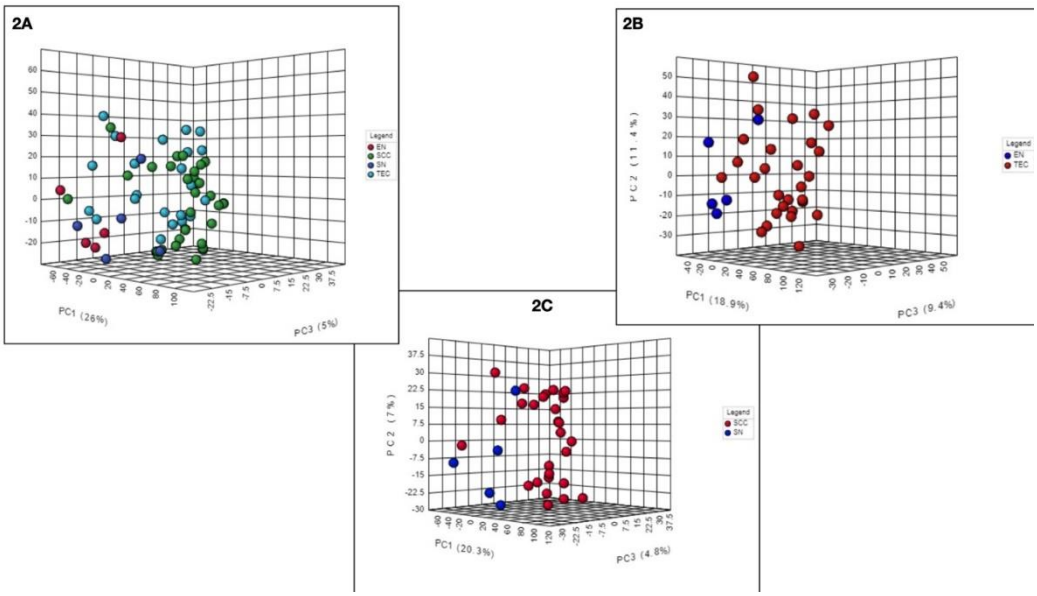

Figure 2: (A) PCA for All 4 conditions (Cancer Epithelial tissue, Cancer Stromal tissue and healthy control epithelial and stromal samples), (B) PCA For Cancer Epithelial tissue samples Vs Normal epithelial tissue samples, (C) PCA showing Variance among Cancer Stromal tissue samples and Normal epithelium tissue samples.

3.2 Downstream Analysis

The PCA results (Figure 2A to 2C) clearly show that the variation is greatest between the cancer stromal tissue and normal epithelial tissue samples (SCC vs SN). Hence, we compared tissue samples from cancer stromal tissue and normal epithelium (SCC versus SN). Because they displayed the second-highest degree of variation in the PCA plots, we also conducted a comparative study between samples of cancer epithelial tissues and normal epithelial tissue (TEC versus TN). SCC and SN samples are well distinguished in Figure 2C, indicating that

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

there is substantial transcriptome diversity between these groups. As a result, we contrasted the gene expression of cancer stromal tissues (SCC vs SN) and cancer epithelial tissues (TEC vs TN) in the downstream analysis with that of normal epithelial tissues. The principal components had improved and the two groups could now be separated from one another when these data were analyzed in a specific scatter plot. We can make some inferences based on the differences that were found between these two groups—Cancer Stromal Tissue vs. Normal Epithelium Tissue [SCC vs SN] samples and Cancer Epithelial Tissues vs. Normal Epithelium Tissue [TEC vs TN] samples. These acquired variations could have further effects on other biological levels, which could be studied at the level of gene regulation.

3.3 Differential Genes Expression Analysis

The examination of the differential gene expression between cancer epithelial tissues and normal epithelial tissue (TEC vs TN) samples was done, and the results were significant (p.adj value 0.1, Fold change >= 1.5). Of them, 46 genes were discovered to be considerably upregulated (p.adj value 0.1, Fold change >= +1.5) in TEC compared to EN, whereas 3060 genes were discovered to be significantly downregulated (p.adj value 0.1, Fold change = -1.5). The analysis of the differential gene expression between the cancer stromal tissues and normal stromal tissues (SCC vs SN) samples looked at 22277 samples significantly (p.adj value 0.1, Fold change >= 1.5). These genes were divided into 62 that were considerably elevated (p.adj value 0.1, Fold change >= +1.5) in SCC compared to SN and 38 that were significantly downregulated (p.adj value 0.1, Fold change = -1.5).

3.4 Clustering and Heat Map revealed variations among Epithelial and Stromal cancer tissues and Normal Control samples

In order to determine whether the differentially expressed significant genes can form distinct clusters of Cancer stromal tissues vs. Normal stromal tissue [SCC vs. SN] samples and Cancer epithelial tissues vs. Normal epithelial tissue [TEC vs. TN] samples based on their gene expression, hierarchical clustering (visualized in the form of dendrograms) (17) was performed. As shown in Figures 3 and 4, respectively, the clustering analysis findings clearly reveal the unique clusters of the group of Cancer stromal tissues vs. Normal stromal tissue [SCC vs SN] samples and Cancer stromal tissues vs. Normal epithelial tissue [TEC vs TN] samples. Heatmap representing the expression pattern of significant genes among Cancer epithelial tissues vs Normal epithelial tissue [TEC vs TN] samples and Cancer stromal tissues vs Normal stromal tissue [SCC vs SN] samples, as shown in Figure 5 and 6 respectively.

ISSN 1533-9211

DOI 10.17605/OSF.IO/DG6TK

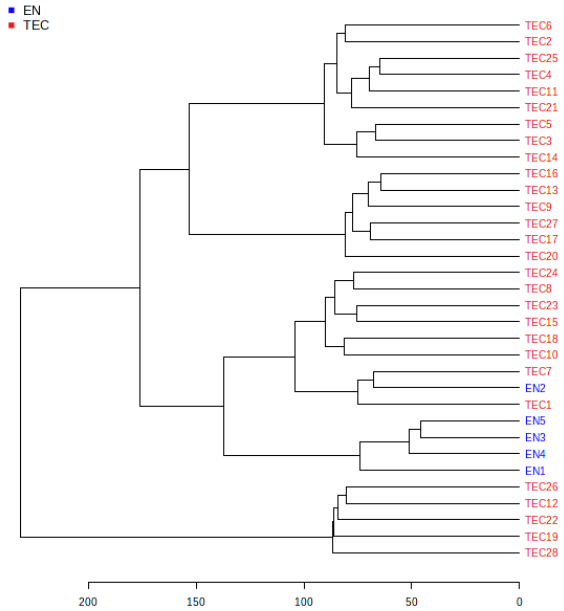

Figure 3: Hierarchical Clustering results as dendrograms. Red text represents the clusters of the diseased (Cancer epithelial tissues) samples and Blue text clusters indicate the Control (Normal epithelial tissue) samples

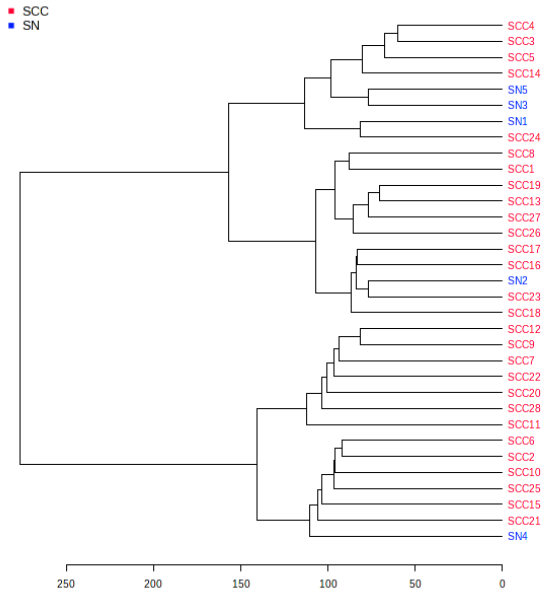

Figure 4: Hierarchical Clustering results as dendrograms. Red text represents the clusters of the diseased (Cancer stromal tissues) samples and blue text clusters indicate the Control (Normal stromal tissue) samples

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

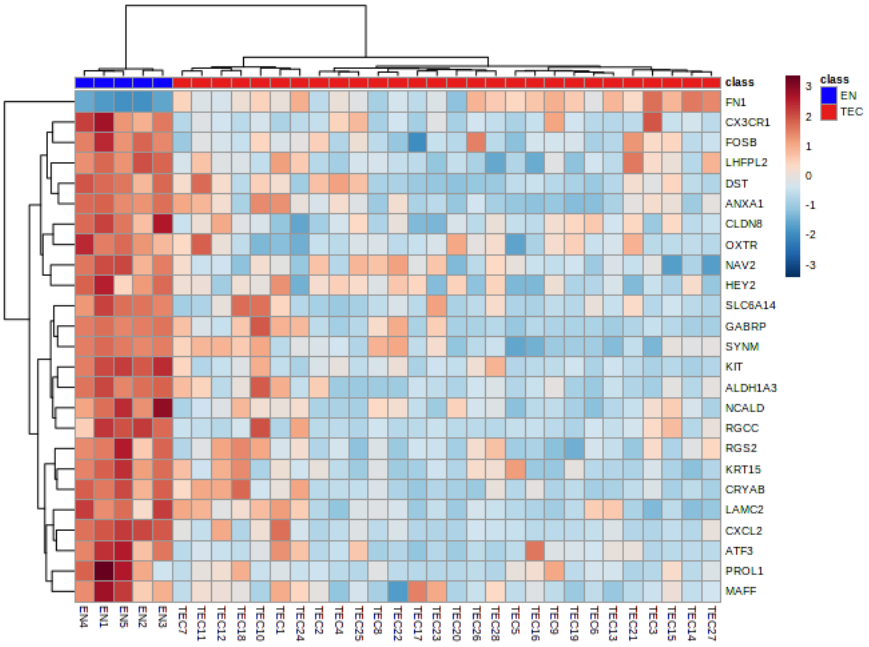

Figure 5: Heat map showing the gene expression profiles of genes with notable differences in expression

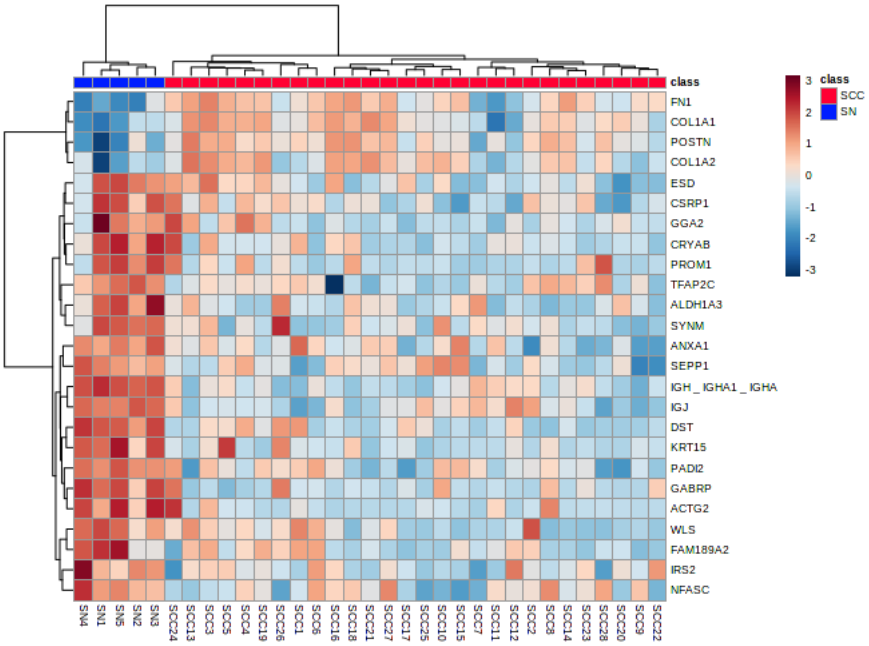

Figure 6: Heat map showing the gene expression profiles of genes with notable differences in expression

It is very much evident from Figure 4 and 5, the clustering patterns formed by the dendrograms, indicating a very clear view of feature selection that could be performed between TEC vs EN

ISSN 1533-9211

DOI 10.17605/OSF.IO/DG6TK

and SCC vs SN groups. Whereas using a feature selection algorithm of the inbuilt metaboanalyst software we scrutinized a heat map of top 25 differentially expressed genes to visualize its expression levels. Our heatmap is clearly able to portray the up-regulated and down regulated genes for the paired groups TEC vs EN and SCC vs SN.

3.5 Pathways involved in the pathogenesis of Epithelial/Stromal cells driven Breast Cancer

The investigation of gene ontologies was done to comprehend the biological significance of pathways. Though different pathways were impacted by the over-expressed and upregulated genes, it is interesting to note that many of these pathways represented biological ones, including the extracellular matrix organization pathway, the protein digestion and absorption pathway, and the platelet derived growth factor binding pathway (Figure 7A). The RNA binding routes, focal adhesion pathways, and SRP dependent co-translational pathways were the biological pathways that were altered as a result of the down-regulated genes (Figure 7B).

To understand the biological importance of pathways, gene ontologies were investigated. It's interesting that, despite the fact that the over-expressed-upregulated genes affected a variety of pathways, many of these pathways represented biological pathways, such as the pathways for the organization of the extracellular matrix, collagen fibrils, and platelet derived growth factor binding (Figure 7C).Cytosolic-large ribosomal routes, cadherin binding pathways, and SRP dependent co-translational processes were the biological pathways that were altered as a result of the down-regulated genes (Figure 7D).

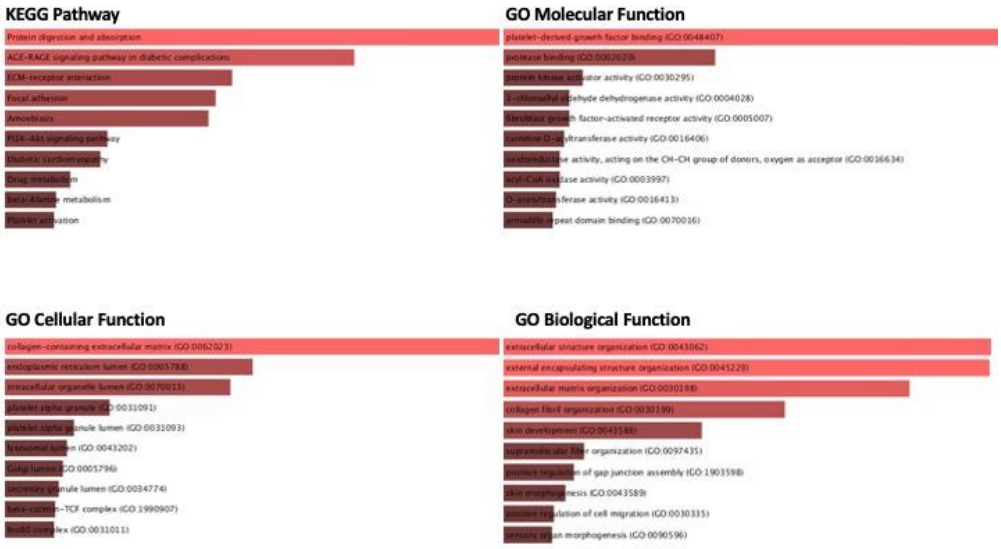

Figure 7A: KEGG pathway and Gene ontology analysis of the Up-regulated genes [TEC vs EN samples]

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

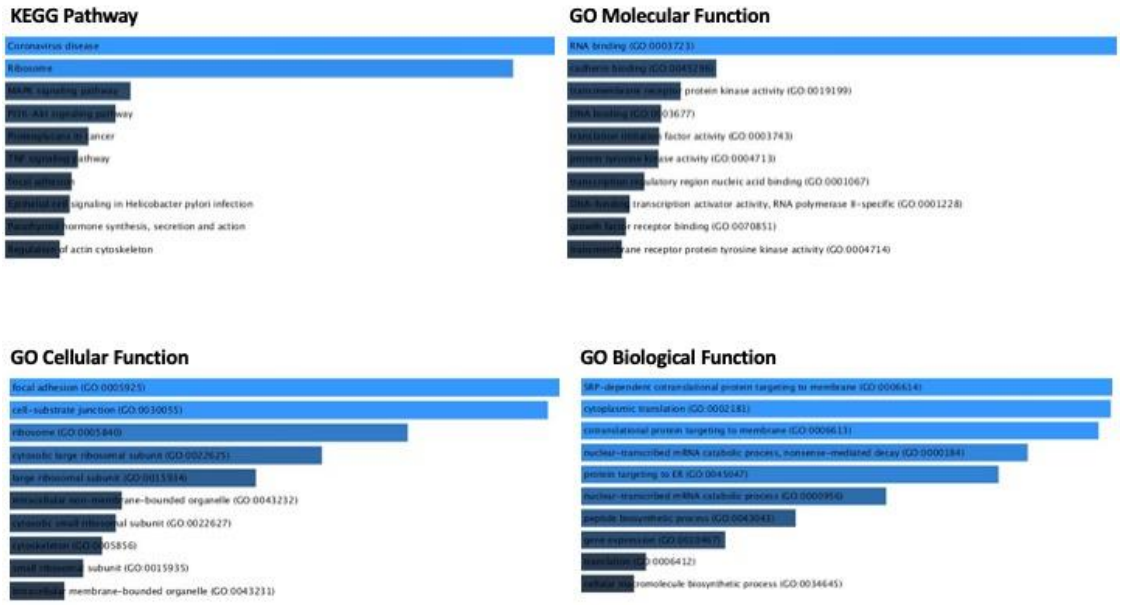

Figure 7B: KEGG pathway and Gene ontology analysis of the Down-regulated genes [TEC vs EN samples]

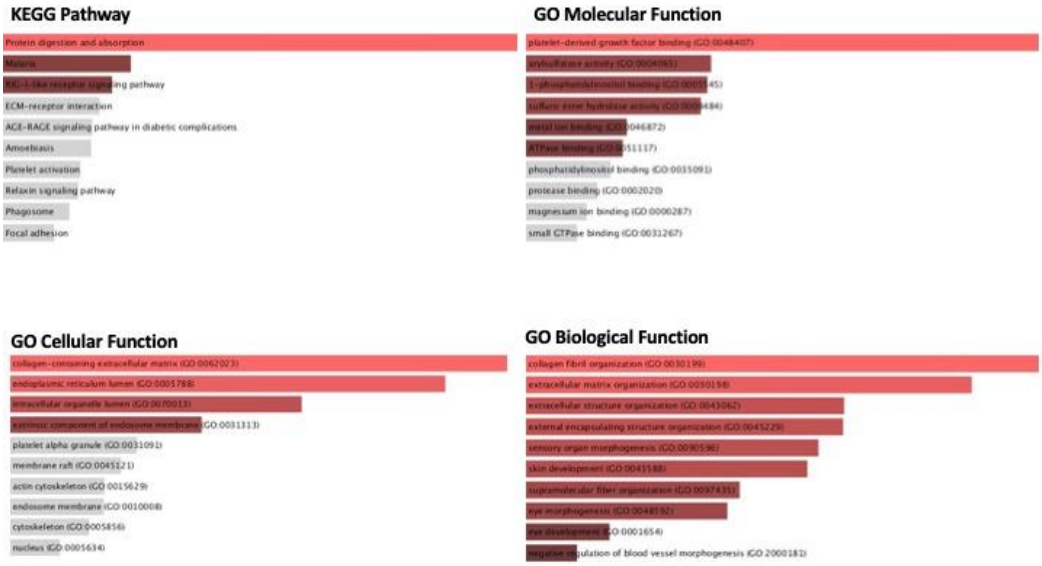

Figure 7C: KEGG pathway and Gene ontology analysis of the Up-regulated genes [TEC vs EN samples]

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

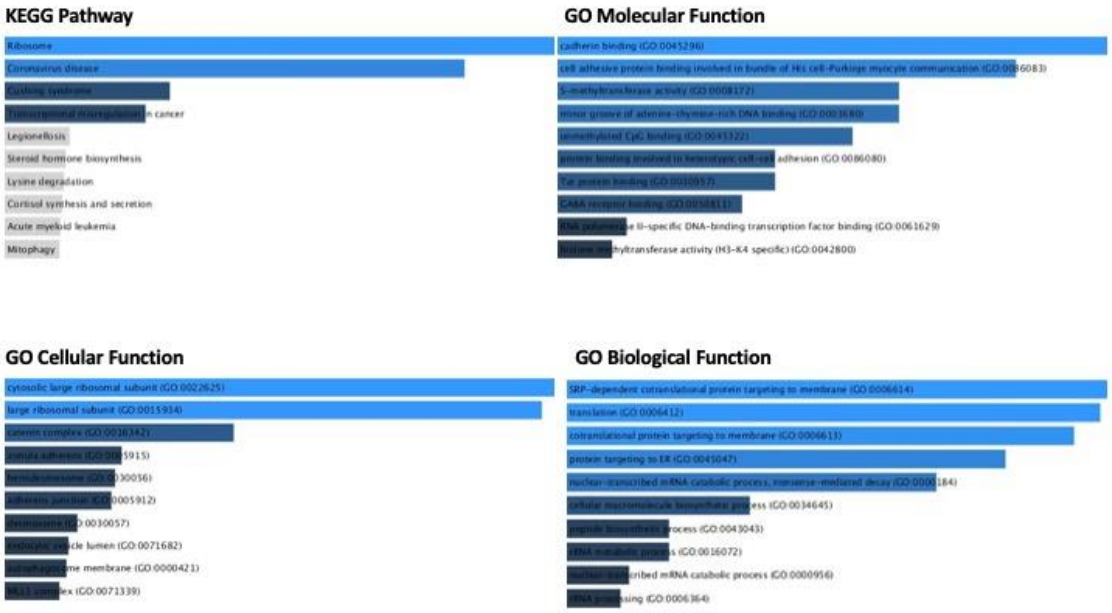

Figure 7D: KEGG pathway and Gene ontology analysis of the Down-regulated genes [SCC vs SN samples]

4. DISCUSSION

Stromal and epithelial cells make up the mammary gland, and they communicate with one another through the extracellular matrix (ECM). Both the induction and promotion of breast cancer can result from disruption of the epithelium-connection (18, 19). Stroma's for the typical mammary gland to form and function correctly, there must be crosstalk between the breast epithelium and stroma. It's interesting to note that the mammary gland exhibits numerous characteristics linked to breast cancer during its developmental cycle (20). Furthermore, a large number of the elements linked to breast cancer are also essential for mammary growth (13, 21). If we have a better understanding of how these components function throughout normal development, we may be able to better understand how cancers begin and grow. In this study, we sought to demonstrate a strong relationship between cancerous tissues from the stroma and epithelium and a number of cellular and molecularly altered biological processes.

In order to identify differences in gene expression between cancer stromal tissues vs. normal stromal tissue [SCC vs SN] samples and cancer stromal tissues vs. normal epithelial tissue [TEC vs TN] samples, RNA-Seq analysis was conducted on RNA-seq samples obtained from both groups. The cancer stromal tissues vs. normal stromal tissues [SCC vs SN] and the cancer stromal tissues vs. normal epithelial tissues [TEC vs TN] samples differ significantly at the gene level, as shown by the samples from both categories forming different clusters. Exploratory data analysis using PCA revealed this. This raises the possibility that there may be a trigger for the development of tumours at the transcriptome level and the progression of

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

metastatic breast cancer. The sequence of events that take place during this transition could be revealed with more study on the subject.

We found that the Cancer epithelial tissues vs. Normal epithelial tissue [TEC vs TN] samples examined 22277 substantially (p.adj value 0.1, Fold change >= 1.5) based on differential gene expression analysis using Welch's T-Test. A total of 46 genes were discovered to be considerably upregulated (p.adj value 0.1, Fold change >= +1.5) in TEC compared to EN, while 3060 genes were found to be significantly downregulated (p.adj value 0.1, Fold change = -1.5). In addition, 22277 significantly different gene expression patterns between cancerous stromal tissues and normal stromal tissues (SCC vs SN) samples were examined (p.adj value 0.1, Fold change >= 1.5). In SCC compared to SN, 62 of the total genes were found to be considerably upregulated (p.adj value 0.1, Fold change >= +1.5), while 38 of them were found to be significantly downregulated (p.adj value 0.1, Fold change = -1.5).

Moreover, such visualization patterns of feature selection such as dendrograms and heatmaps help one to select proper machine learning algorithm to extract features. These visualization patterns, could later be used to work on breast cancer sub-features selection such as classification based on gender, age, race, ethnicity and pathophysiology.

The development of tumours in epithelial/stromal tissues may have been influenced by possible dysregulation, according to a gene ontology analysis using Enrichr based on significant gene sets. This analysis revealed obvious involvement of genes in extracellular matrix organization pathways, protein digestion and absorption pathways, platelet derived growth factor binding pathways, RNA binding pathways, focal adhesion pathways, and SRP dependent co-translational pathways.

This study shows that as per our GO Cellular Component 2021 pathway Analysis, the most affected pathway is the collagen-containing extracellular matrix pathway. The associated genes with this pathway are COL11A1, COL1A1, COL1A2, COL3A1, COL5A1 and COL5A2 [on the basis of analysis done on the Enrichr software]as the key features that may substantially contribute to metastasis of breast cancer from epithelial cells to stromal cells in the mammary glands. Studies have shown that Collagen expression is increased during breast cancer development. Another evidence that moderate collagen expression is adequate to improve cancer cell stemness comes from the fact that overexpression of collagen in tumour cells had minimal impact on the development of the tumour microenvironment. Moreover, studies have demonstrated that collagen accelerates the development of breast cancer by increasing the stemness and anoikis resistance of cancer cells (22). Our study depicts the collagen genes such as COL11A1, COL1A1, COL1A2, COL3A1, COL5A1 and COL5A2 that are highly up-regulated might be the possible cause of over production of collagen. According to recent studies, collagens can also affect the phenotypic and operation of a variety of immune cells that infiltrate tumours, including tumor-associated macrophages (TAMs) and T cells. Cancer cells' migratory and proliferative rate may be accelerated by collagen (23). Although collagen have the ability to keep the immune cells out of tumours and proliferation process, studies have also suggested that the loss or overly gain of collagen can allow tumors to grow more rapidly (23).

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

Anoikis is a type of programmed cell death that happens when a cell separates from the appropriate extracellular matrix and interferes with integrin ligation (24). It is a crucial mechanism for stopping the proliferation of dysplastic cells or their attachment to the wrong kind of matrix. Anoikis in cancer cells retards metastasis of cells to other sites. The survival of cancer cells during cancer spread depends on their resistance to detachment-induced anoikis, which is more prevalent in tumour initiating cells (25). Collagen, according to studies, accelerates the spread of breast cancer by increasing the stemness and anoikis resistance of cancer cells (26). There is still some question about whether gene dysregulation caused by collagen enhances anoikis resistance.

Also, the over-expression of collagen producing genes in epithelial cancer tissue is an important biomarker. However, is it a result of a mutated gene or a signaling pathway is a crucial benchmark to cross? One of the main drivers of cancer cell behavior and a factor in how cancer cells interact with ECM elements is the heterogeneity of mutant genes (26, 27). The circumstances for collagen in the tumour matrix are also changed by the mutation of oncogenes, which are primarily split into tumour suppressor genes and proto-oncogenes.

Mammary gland branching morphogenesis is significantly influenced by interactions between the epithelium and ECM. When ECM-regulating factors are inhibited or deleted, TEB production and ductal invasion are hampered. We are also aware that expression, by encouraging cancer cell invasion and stemness, may aid in the colonization and spread of cancer (27). Particularly, it is thought that the primary source of ECM protein in cancer tissue is cancer-associated fibroblasts (28). It has been discovered that fibroblasts considerably express collagen, and that it concentrates in the focal adhesion (29, 30). Yet further research is needed to determine what causes the precise up-regulated and down-regulated genes to target the collagen ECM pathway and collagen synthesis.

It has been established that interstitial collagen and BM collagen play a part in the growth of breast tumours (31). In addition, it should be emphasized that in this study, 6 genes—COL11A1, COL1A1, COL1A2, COL3A1, COL5A1, and COL5A2—were identified as crucial critical elements that may help breast cancer spread from epithelial cells to stromal cells in the mammary glands. This study was able to pinpoint a number of transcriptional genes whose function in the relationship between breast cancer metastasis and association remains mostly unknown. Further research into how they function might identify important indicators or pharmacological targets that could be used to treat breast cancer in conjunction with mammary epithelial and stromal cells. In order to detect early signs of cancer metastasis, these genes can be examined in breast cancer patients who may be subject to screening for the disease.

Conclusively, our study is able to reveal some of the potential significant differences in gene expression between Cancer epithelial tissues vs Normal epithelial tissue [TEC vs TN] samples and Cancer stromal tissues vs Normal stromal tissue [SCC vs SN] samples. Interestingly, we also identified the affected biological pathways for both cancer stromal tissues vs. normal stromal tissues [SCC vs. SN] samples and cancer epithelial tissues vs. normal epithelial tissue [TEC vs. TN] samples as a result of the up-regulated and down-regulated genes. This most certainly provides a crucial hint about the cause of the deadly metastatic cancer condition.

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

5. FUTURE DIRECTIONS

To determine the precise impact of breast cancer on genes like COL11A1, COL1A1, COL1A2, COL3A1, COL5A1, and COL5A2, much study will be needed in the future. To back up the aforementioned conclusions, more thorough investigation is required. It would be interesting to learn more about how the pathogenesis of breast cancer affects the extracellular matrix organization pathways, protein digestion and absorption pathways, platelet derived growth factor binding pathways, RNA binding pathways, focal adhesion pathways, and SRP dependent co-translational pathways as part of a future study.

Moreover, in order to elucidate dendrograms and Heat-maps towards the potential of using their visualization patterns, as an important tool for precise feature selection. Such selected features could be then used further in order to demonstrate important gene level differences between different experimental groups.

References

1. Houghton, S. C., & Hankinson, S. E. (2021). Cancer Progress and Priorities: Breast Cancer. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 30(5), 822–844. https://doi.org/10.1158/1055-9965.EPI-20-1193
2. Kumar N, Patni P, Agarwal A, Khan MA, Parashar N. Prevalence of molecular subtypes of invasive breast cancer: a retrospective study. Med J Armed Forces India. 2015;71(3):254–8.
3. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol. 2012;196(4):395–406.
4. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.
5. Iyengar P, Espina V, Williams TW, Lin Y, Berry D, Jelicks LA, Lee H, Temple K, Graves R, Pollard J, et al. Adipocyte-derived collagen VI affects early mammary tumor progression in vivo, demonstrating a critical interaction in the tumor/stroma microenvironment. J Clin Invest. 2005;115(5):1163–76.
6. Provenzano PP, Inman DR, Eliceiri KW, Knittel JG, Yan L, Rueden CT, White JG, Keely PJ. Collagen density promotes mammary tumor initiation and progression. BMC Med. 2008;6:11.
7. Shields MA, Dangi-Garimella S, Krantz SB, Bentrem DJ, Munshi HG. Pancreatic cancer cells respond to type I collagen by inducing snail expression to promote membrane type 1 matrix metalloproteinasedependent collagen invasion. J Biol Chem. 2011;286(12):10495–504. 8. Condeelis J, Segall JE. Intravital imaging of cell movement in tumours. Nat Rev Cancer. 2003;3(12):921–30.
8. Provenzano PP, Inman DR, Eliceiri KW, Keely PJ. Matrix density-induced mechanoregulation of breast cell phenotype, signaling and gene expression through a FAK-ERK linkage. Oncogene. 2009;28(49):4326–43.
9. Mol Pathol. 2000 Apr; 53(2): 64–68. doi: 10.1136/mp.53.2.64
10. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356(3):227–36.
11. Guo YP, Martin LJ, Hanna W, Banerjee D, Miller N, Fishell E, Khokha R, Boyd NF. Growth factors and stromal matrix proteins associated with mammographic densities. Cancer Epidemiol Biomark Prev. 2001;10(3):243–8.

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

12. Kalluri, R., & Zeisberg, M. (2006). Fibroblasts in cancer. Nature reviews. Cancer, 6(5), 392–401. https://doi.org/10.1038/nrc1877.
13. Myllyharju J, Kivirikko KI. Collagens, modifying enzymes and their mutations in humans, flies and worms. Trends Genet. 2004;20(1):33–43.
14. Hellewell AL, Adams JC. Insider trading: extracellular matrix proteins and their non-canonical intracellular roles. Bioessays. 2016;38(1):77–88.
15. J. Lever, M. Krzywinski, N. Altman, Points of Significance: Principal component analy-sis. Nature Methods, 14(7), 641–642 (2017). doi:10.1038/nmeth.4346
16. Fu Q, Hoijtink H, Moerbeek M. Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. Behav Res Methods. 2021 Feb;53(1):139–152. doi: 10.3758/s13428-020-01408-1. PMID: 32632740; PMCID: PMC7880954.
17. Kavuri VC, Liu H. Hierarchical clustering method to improve transrectal ultrasound-guided diffuse optical tomography for prostate cancer imaging. Acad Radiol. 2014 Feb;21(2):250–62. doi: 10.1016/j.acra.2013.11.003. PMID: 24439338; PMCID: PMC4562019.
18. Banyard J, Bao L, Zetter BR. Type XXIII collagen, a new transmembrane collagen identified in metastatic tumor cells. J Biol Chem. 2003;278(23):20989–94.
19. Hashimoto T, Wakabayashi T, Watanabe A, Kowa H, Hosoda R, Nakamura A, Kanazawa I, Arai T, Takio K, Mann DM, et al. CLAC: a novel Alzheimer amyloid plaque component derived from a transmembrane precursor, CLAC-P/collagen type XXV. EMBO J. 2002;21(7):1524–34.
20. Hagg P, Rehn M, Huhtala P, Vaisanen T, Tamminen M, Pihlajaniemi T. Type XIII collagen is identified as a plasma membrane protein. J Biol Chem. 1998; 273(25):15590–7.
21. Maatta M, Vaisanen T, Vaisanen MR, Pihlajaniemi T, Tervo T. Altered expression of type XIII collagen in keratoconus and scarred human cornea: increased expression in scarred cornea is associated with myofibroblast transformation. Cornea. 2006;25(4):448–53.
22. Taddei, M. L., Giannoni, E., Fiaschi, T., & Chiarugi, P. (2012). Anoikis: an emerging hallmark in health and diseases. The Journal of pathology, 226(2), 380–393. https://doi.org/10.1002/path.3000
23. Nykvist P, Tu H, Ivaska J, Kapyla J, Pihlajaniemi T, Heino J. Distinct recognition of collagen subtypes by alpha(1)beta(1) and alpha(2)beta(1) integrins. Alpha(1)beta(1) mediates cell adhesion to type XIII collagen. J Biol Chem. 2000;275(11):8255–61.
24. Vaisanen MR, Vaisanen T, Pihlajaniemi T. The shed ectodomain of type XIII collagen affects cell behaviour in a matrix-dependent manner. Biochem J. 2004;380(Pt 3):685–93.
25. Snellman A, Keranen MR, Hagg PO, Lamberg A, Hiltunen JK, Kivirikko KI, Pihlajaniemi T. Type XIII collagen forms homotrimers with three triple helical collagenous domains and its association into disulfide-bonded trimers is enhanced by prolyl 4-hydroxylase. J Biol Chem. 2000;275(12): 8936–44.
26. Snellman A, Tu H, Vaisanen T, Kvist AP, Huhtala P, Pihlajaniemi T. A short sequence in the N-terminal region is required for the trimerization of type XIII collagen and is conserved in other collagenous transmembrane proteins. EMBO J. 2000;19(19):5051-9
27. Vaisanen T, Vaisanen MR, Autio-Harmainen H, Pihlajaniemi T. Type XIII collagen expression is induced during malignant transformation in various epithelial and mesenchymal tumours. J Pathol. 2005;207(3):324–35.
28. Miyake M, Hori S, Morizawa Y, Tatsumi Y, Toritsuka M, Ohnishi S, Shimada K, Furuya H, Khadka VS, Deng Y, et al. Collagen type IV alpha 1 (COL4A1) and collagen type XIII alpha 1 (COL13A1) produced in

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

cancer cells promote tumor budding at the invasion front in human urothelial carcinoma of the bladder. Oncotarget. 2017;8(22):36099–114.

29. Plantefaber LC, Hynes RO. Changes in integrin receptors on oncogenically transformed cells. Cell. 1989;56(2):281–90. 27. Cosgrove D, Rodgers K, Meehan D, Miller C, Bovard K, Gilroy A, Gardner H, Kotelianski V, Gotwals P, Amatucci A, et al. Integrin alpha1beta1 and transforming growth factor-beta1 play distinct roles in alport glomerular pathogenesis and serve as dual targets for metabolic therapy. Am J Pathol. 2000;157(5):1649–59. 28. Howe AK, Aplin AE, Juliano RL. Anchorage-dependent ERK signaling-- mechanisms and consequences. Curr Opin Genet Dev. 2002;12(1):30–5.
30. George EL, Georges-Labouesse EN, Patel-King RS, Rayburn H, Hynes RO. Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin. Development. 1993;119(4):1079–91.
31. Hynes RO. Targeted mutations in cell adhesion genes: what have we learned from them? Dev Biol. 1996;180(2):402–12.