**Co-expression network analysis of Breast Cancer using RNA-Sequencing (RNA-Seq) dataset**

**Abstract**

Gene co-expression networks have been used for assigning functions to the unknown genes and finding their biological significance. It has been reported in many studies that in cancers there are alterations in the tightly regulated gene regulatory networks which cause the abnormal proliferation of the cancer cells. Following these observations, we used the co-expressed networks of cancer and normal datasets from breast tissue to detect the altered networks. In this project, we compared the co-expression networks of breast tissue in normal and cancer conditions. We used a total of 80 RNA-Seq samples to build the co-expression networks in cancer and normal conditions. By using differential network analysis, we found that despite having similar genes, the networks differ completely in terms of the gene-pair connections. These results show that in breast tissue how differently genes interact in cancer compared to normal cells. We also found some interesting gene clusters which are enriched in terms including cell adhesion and development, and mammary cancer.

**Introduction**

Breast cancer is one of the life-threatening cancers among the women in the US coming after lung cancer. According to the recent statistics, 1 in every 8 US women develops breast cancer in her lifetime. The ratio of men having breast cancer is around 1 in every 1000 men. This signifies that the risk of a woman having breast cancer is much more than a man. It has been estimated that in the year 2017, 250 thousand new cases of invasive breast cancer are expected to be diagnosed in women in the US [1]. These statistics clearly states that how important it is to research the mechanisms behind breast cancer development and progression. There has been research going on in this area and researchers were able to classify breast cancer based on their gene expression profile into four main intrinsic molecular subtypes of breast cancer known as luminal A, luminal B, HER2-enriched [HER2E] and basal-like [2].

Around a decade ago, the cancer research took a new turn when the next-generation sequencing (NGS) techniques were introduced. NGS techniques provide researchers a tool to study unknown transcript isoforms which were not captured earlier by microarray techniques. An analysis similar to that performed using microarray can also be performed using RNA-Seq, but with more confidence and on a larger family of RNA molecules. Apart from being cost-effective these techniques also provide high confidence in the analysis. A lot of studies have been published in order to study the gene expression profiles in the normal and the cancer cells in order to find the possible drug targets. Most of these studies were done on the gene level which is also called as differential expression analysis [3–5]. Although these studies provide us an insight into the importance of the various gene which is important for the development of the breast cancer including BRCA1 and BRCA2 [6] but the inherent organization of the genes has largely remained unexplored. Co-expression network analysis will help us to reveal global expression patterns of genes [7]. Co-expression network analysis tells us about the varying degree of co-expression levels between the pairs of genes which can be used to identify the regulations behind the cancer development. It has also extensively been used for discovery of disease marker genes and inferring gene regulatory networks. Clustering techniques are also used to identifying the biologically relevant modules of genes in the networks.

In this study, we want to compare the co-expression networks of the normal and cancer dataset. Using the co-expression networks, we want to identify the co-expressed gene sets or clusters which play important role in the cancer development and progression. In order to study this, we built co-expression networks for the normal and cancer samples using different correlation methods. There are two major types of correlation methods namely Pearson (Linear) and Spearman (Non-Linear). For our study, we first assessed these methods by comparing the networks build from these two methods. After that, we used the better method to compare the normal and cancer networks using the iGraph package in R [8]. Using SPICi [9] and GREAT [10], we found two interesting gene clusters which appear to be biologically important for the development and progression of breast cancer.

**Methods**

**Data processing**

Cancer and the normal breast RNA-Seq data was downloaded from The Cancer Genome Atlas (TCGA) database [11]. The downloaded data consists of the gene expression of the samples in FPKM (Fragment per kilo lengths) units for each sample. We downloaded a total of 80 samples 40 for each condition. In the first step, we combined the expression profile of all cancer and the normal samples and created gene expression matrix consisting of rows as genes and columns as samples. After that, we converted the ensemble ids to Gene symbol by using Biomart [12]. Lastly, we filtered out all the genes which had FPKM values less than 1. In the end, we got a total of 15,859 genes for our analysis.

**Co-expression network construction**

For building the co-expression network we followed a simple approach i.e. usage of correlation functions. We calculated the correlation values for each gene using the gene-expression matrix which we created earlier. In previous studies, Pearson correlation has been used for creating Co-expression networks [13–15]. As mentioned earlier, we want to compare these two methods, so we used both the Pearson and Spearman correlation coefficients to calculate the co-expression correlation between all the genes. After that, the correlation cutoff for the selection of the edges is calculated by plotting the network density values at different correlation cutoffs.

**Significance of largest connected sub-network**

In order to check the significance of the co-expression network that we get in the previous step, we compared it with random networks having the same number of nodes and edges. For building the random networks we used the “Erdos-Renyi” model in iGraph. We compared various network properties of the networks including the clustering coefficient and average path length. We used Wilcox signed rank test to show that our network is significantly different than the random network.

**Differential network analysis**

The co-expression networks developed for different conditions can be compared by carrying out simple operations on their edges and nodes. We carry out differential network analysis by calculating the number of genes and edges that are there in a specific condition. In our case, we calculated the condition-specific genes or nodes and edges for cancer and normal datasets. For example, if we take networks of two conditions, C and N, having nodes c and n nodes, we can define differential nodes:

DC = c – n

DN = n – c

where DC is the number of nodes present only in condition C network and DN is the number of nodes present only in condition N network. Similarly, we can calculate the condition specific edges between two genes. We used the iGraph package in R [8] for carrying out the differential network analysis.

**Network Clustering**

For finding modules in the breast cancer co-expression network we used SPICi (Speed and Performance in Clustering) [9]. This tool uses a greedy heuristic approach to find clusters in the network. In this algorithm, only one cluster is built at a time and it uses a seed pair (proteins or genes) to build the cluster. The seed pair is selected based on the weight of the degree. After that, the cluster is expanded by adding the adjacent nodes till the cluster density is greater than the threshold. The procedure is repeated until all the nodes in the network are used. This approach makes it very fast with very low memory requirement as compared with other clustering methods including MCL [16]. For our analysis, we used the cluster density threshold as 0.5 which is recommended by the authors of the tool. We also restricted the clusters with low genes by using the minimum cluster size as 30.

**Gene ontology analysis**

To find the biological processes linked to the different gene clusters found by using SPICi, gene ontology analysis was performed using GREAT and Revigo [10][17]. We used hg19 human reference genome for getting the GO terms. The significance of the ontology terms was calculated by using hypergeometric test. We filter out the terms by taking a threshold of 0.05 for the FDR level. The ontology terms are visualized by using treemap package in R. The size of the element in the treemap is represented by the fold-change of the terms and the color represents the significance of the term (-log10(FDR)).

**Results**

**Correlation networks**

We used both the Pearson and Spearman correlation coefficient for creating the co-expression networks using the gene expression dataset for both normal and cancer samples. As mentioned earlier, we filtered the edges based on a correlation threshold. In our case, we calculated the degree distribution of the various thresholds and from that, we found that the degree distribution stabilizes at a threshold of 0.55 (Figure 1). In order to make our network stronger, we selected a threshold of 0.7 which is a bit higher. After filtering the edges, we calculated the properties of the networks which is shown in Table 1.

*Figure 1: Degree Distribution V/S Correlation Threshold.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Cancer Samples** | | **Normal Samples** | |
| **Pearson** | **Spearman** | **Pearson** | **Spearman** |
| **Nodes** | 14162 | 11334 | 13668 | 13294 |
| **Edges** | 1063535 | 161255 | 4789107 | 5181987 |
| **Density** | 0.010606 | 0.0025108 | 0.05127 | 0.0586472 |
| **Diameter** | 14 | 17 | 11 | 12 |
| **Clustering Coefficient** | 0.5112 | 0.3595 | 0.614085 | 0.605653 |
| **Shortest Path** | 4.515 | 5.371 | 3.279 | 3.26862 |
| **Sub-Networks** | 62 | 228 | 11 | 16 |

*Table 1: Network Properties of the Pearson and Spearman co-expression networks.*

Table 1 shows that the co-expression network consists of a large number of sub-networks. Most of these sub-networks consist of a small number of nodes. So, for further analysis, we used the largest connected sub-network. The various network properties of the largest connected sub-network are shown in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Cancer Samples** | | **Normal Samples** | |
| **Pearson** | **Spearman** | **Pearson** | **Spearman** |
| **Nodes** | 14028 | 10805 | 13645 | 13261 |
| **Edges** | 1063460 | 160839 | 4789094 | 5181969 |
| **Density** | 0.0108091 | 0.002755 | 0.0514479 | 0.058939 |
| **Diameter** | 14 | 17 | 11 | 12 |
| **Clustering Coefficient** | 0.51142 | 0.35883 | 0.61422 | 0.605793 |
| **Shortest Path** | 4.515 | 5.3717 | 3.2793 | 3.268 |

*Table 2: Network Properties of the Pearson and Spearman largest connected co-expression networks.*

**Significance of largest connected sub-network**

After that, we want to check whether our network is significant for the specific conditions and representing the developmental process. For this, we compared our network with 10 random networks generated by using “Erdos-Renyi” model in iGraph. We calculated the network properties of these random networks including the clustering coefficient and average path length. By using Wilcox-signed rank test, it was found that the average path length (P-Value = 0.0004883) and clustering coefficient (P-Value=0.0009766) of our networks are significantly different than the random network.

**Comparison of Pearson and Spearman co-expression networks**

One of our aims in this project is to compare the co-expression networks created from Pearson and Spearman correlation coefficient. For this, we compared the nodes and edges of these two networks in the two different conditions. The results are shown in Table 3. It was seen that the Pearson co-expression network have more differential nodes and edges as compared to Spearman network in cancer condition. In normal condition, both of the networks shows a similar number of edges and nodes (compared to the cancer network). From these results, we can say that Pearson correlation coefficient is good for capturing the co-expression relationship. So, for our further analysis, we used Pearson co-expression network.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Cancer Samples** | | **Normal Samples** | |
| **Pearson** | **Spearman** | **Pearson** | **Spearman** |
| **Differential Nodes** | 3748 | 525 | 517 | 133 |
| **% Differential Nodes** | 26.71 % | 4.85 % | 3.79 % | 1.01 % |
| **Differential Edges** | 994862 | 92241 | 953814 | 1346689 |
| **% Differential Edges** | 93.54 % | 57.34 % | 19.91 % | 25.98 % |

*Table 3: Differential nodes and edges between the Pearson and Spearman networks in cancer and normal samples.*

**Comparison of normal and cancer co-expression networks**

In this case, we compared the normal and breast cancer co-expression networks by calculating the differential nodes and genes as mentioned in the methods section. The results are shown in table 4.

|  |  |  |
| --- | --- | --- |
|  | **Cancer Samples** | **Normal Samples** |
| **Total Nodes** | 14028 | 13645 |
| **Differential Nodes** | 2278 | 1895 |
| **% Differential Nodes** | 16.21 % | 13.88 % |
| **Total Edges** | 1063460 | 4789094 |
| **Differential Edges** | 928753 | 4654387 |
| **% Differential Edges** | 87.33 % | 97.18 % |

*Table 4: Differential nodes and edges between cancer and normal samples.*

Although, the number of differential genes between the normal and cancer networks is not very big (16.21% in cancer network and 13.88% in the normal network) but the number of edges or gene pairs in these networks is completely different. This shows that just not the expression of one gene but the whole gene clusters work or express together.

**Network Clustering**

We used SPICi network clustering tool for getting the cluster of genes from our condition-specific networks. We got a total of 46 and 41 clusters of cancer and the normal co-expression networks respectively. The clusters were sorted and numbered based on their sizes. After that, we used GREAT to analyze these clusters based on the ontology terms (GO Biological Processes and Osborne Annotated Disease Ontology) related to the clusters. We found that cluster 4 of the breast cancer network shows enrichment of immune GO Biological process (Figure 1) terms which mean that the genes in this cluster are related providing immunity to the body. We checked and found that the expression of these genes is very low in breast cancer dataset. We also found enrichment of cell adhesion and breast cancer terms in cluster 14 (Figure 2 and 3). The expression of these genes is also very high in the RNA-Seq dataset. From these results, we can conclude that these two clusters are really important for the progression and development of breast cancer.

*Figure 2: GO Biological Process enrichment for Cluster 4 of breast cancer network.*



*Figure 3: GO Biological Process enrichment for Cluster 14 of breast cancer network.*



*Figure 4: Osborne Annotated Disease Ontology term enrichment for Cluster 14 of breast cancer network.*

**Conclusions**

By carrying out the co-expression network analysis, we have explored the area of the transcriptional organization in the development and progression of breast cancer. There are several important things that we found after analyzing the co-expression networks. Firstly, the networks that we developed from the Pearson and Spearman correlation methods are quite different which is expected. From our differential network analysis, it was found that the Pearson network is bigger than the Spearman one and also have more unique edges. So, from these results, we found that the breast cancer gene expression dataset follows the linear distribution. We also found that in many studies, researchers used Pearson correlation for carrying out co-expression analysis {reference}. So, for our further analysis, we used Pearson co-expression network.

After that, we compared the network properties of normal and cancer co-expression networks to get an insight. Although, the two networks share almost 80% of the genes but have completely different interactions between those nodes. We found that these networks have almost ~88% differential edges which give us an insight into how the same set of genes work differentially in normal and cancer conditions. To further study these networks, we divided the networks and got the highly-connected parts or modules by using SPICi. After that, we analyzed these modules by carrying out the GO terms enrichment using GREAT. We found interesting clusters which have cancer cell development and differentiation enriched terms and also have high expression in the breast cancer samples. We also found a cluster which is enriched with the immune terms and have low expression in the breast cancer samples. This study has given an insight into the interaction and co-expression of genes in breast cancer. Further research in this area will definitely help us in unfolding mechanism behind breast cancer development.

**References**

[1] U.S. Breast Cancer Statistics | Breastcancer.org, (n.d.). http://www.breastcancer.org/symptoms/understand\_bc/statistics (accessed May 3, 2017).

[2] A. Prat, B. Adamo, M.C.U. Cheang, C.K. Anders, L.A. Carey, C.M. Perou, Molecular characterization of basal-like and non-basal-like triple-negative breast cancer., Oncologist. 18 (2013) 123–33. doi:10.1634/theoncologist.2012-0397.

[3] J.M. Grunda, A.D. Steg, Q. He, M.R. Steciuk, S. Byan-Parker, M.R. Johnson, W.E. Grizzle, Differential expression of breast cancer-associated genes between stage- and age-matched tumor specimens from African- and Caucasian-American Women diagnosed with breast cancer, BMC Res. Notes. 5 (2012) 248. doi:10.1186/1756-0500-5-248.

[4] J. Lane, T.A. Martin, C. McGuigan, M.D. Mason, W.G. Jiang, The differential expression of hCNT1 and hENT1 i n breast cancer and the possible impact on breast cancer therapy., J. Exp. Ther. Oncol. 8 (2010) 203–10. http://www.ncbi.nlm.nih.gov/pubmed/20734919 (accessed May 4, 2017).

[5] F. Peng, Y. Zhang, R. Wang, W. Zhou, Z. Zhao, H. Liang, L. Qi, W. Zhao, H. Wang, C. Wang, Z. Guo, Y. Gu, Identification of differentially expressed miRNAs in individual breast cancer patient and application in personalized medicine, Oncogenesis. 5 (2016) e194. doi:10.1038/oncsis.2016.4.

[6] J.M. Grunda, A.D. Steg, Q. He, M.R. Steciuk, S. Byan-Parker, M.R. Johnson, W.E. Grizzle, Differential expression of breast cancer-associated genes between stage- and age-matched tumor specimens from African- and Caucasian-American Women diagnosed with breast cancer., BMC Res. Notes. 5 (2012) 248. doi:10.1186/1756-0500-5-248.

[7] S. Van Dam, U.V.~ Osa, A. Van Der Graaf, L. Franke, J.~ Ao, P. De Magalh, Gene co-expression analysis for functional classification and gene–disease predictions, (n.d.). doi:10.1093/bib/bbw139.

[8] igraph R package, (n.d.). http://igraph.org/r/ (accessed May 4, 2017).

[9] P. Jiang, M. Singh, SPICi: a fast clustering algorithm for large biological networks, Bioinformatics. 26 (2010) 1105–1111. doi:10.1093/bioinformatics/btq078.

[10] C.Y. McLean, D. Bristor, M. Hiller, S.L. Clarke, B.T. Schaar, C.B. Lowe, A.M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions., Nat. Biotechnol. 28 (2010) 495–501. doi:10.1038/nbt.1630.

[11] NIH, Home - The Cancer Genome Atlas - Cancer Genome - TCGA, (n.d.). https://cancergenome.nih.gov/ (accessed May 4, 2017).

[12] J.M. Guberman, J. Ai, O. Arnaiz, J. Baran, A. Blake, R. Baldock, C. Chelala, D. Croft, A. Cros, R.J. Cutts, A. Di Génova, S. Forbes, T. Fujisawa, E. Gadaleta, D.M. Goodstein, G. Gundem, B. Haggarty, S. Haider, M. Hall, T. Harris, R. Haw, S. Hu, S. Hubbard, J. Hsu, V. Iyer, P. Jones, T. Katayama, R. Kinsella, L. Kong, D. Lawson, Y. Liang, N. Lopez-Bigas, J. Luo, M. Lush, J. Mason, F. Moreews, N. Ndegwa, D. Oakley, C. Perez-Llamas, M. Primig, E. Rivkin, S. Rosanoff, R. Shepherd, R. Simon, B. Skarnes, D. Smedley, L. Sperling, W. Spooner, P. Stevenson, K. Stone, J. Teague, J. Wang, J. Wang, B. Whitty, D.T. Wong, M. Wong-Erasmus, L. Yao, K. Youens-Clark, C. Yung, J. Zhang, A. Kasprzyk, BioMart Central Portal: an open database network for the biological community., Database (Oxford). 2011 (2011) bar041. doi:10.1093/database/bar041.

[13] S. Bergmann, J. Ihmels, N. Barkai, Similarities and Differences in Genome-Wide Expression Data of Six Organisms, PLoS Biol. 2 (2003) e9. doi:10.1371/journal.pbio.0020009.

[14] S.L. Carter, C.M. Brechbuhler, M. Griffin, A.T. Bond, Gene co-expression network topology provides a framework for molecular characterization of cellular state, Bioinformatics. 20 (2004) 2242–2250. doi:10.1093/bioinformatics/bth234.

[15] B. Zhang, S. Horvath, A General Framework for Weighted Gene Co-Expression Network Analysis, Stat. Appl. Genet. Mol. Biol. 4 (2005) Article17. doi:10.2202/1544-6115.1128.

[16] S. van Dongen, C. Abreu-Goodger, Using MCL to Extract Clusters from Networks, in: Methods Mol. Biol., 2012: pp. 281–295. doi:10.1007/978-1-61779-361-5\_15.

[17] F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms, PLoS One. 6 (2011) e21800. doi:10.1371/journal.pone.0021800.