Digital Epidemiology and Precision Medicine

A network-based approach to unveiling crucial nodes for cell phenotypes

December 19, 2021

Report by:

Iliyas Bektas

1 Abstract

In complex network analysis, switch genes are hubs that are not only characterized by a marked negative correlation with their first nearest neighbors, but also distinguished by an unusual pattern of intra- and inter-module connections that gives them a crucial topological role, interestingly mirrored by the evidence of their clinic-biological relevance. In fact, the switch genes were found in all cancers studied by Paci et al., 2017 and they encompass protein coding genes and non-coding RNAs, recovering many known key cancer players but also many new potential biomarkers not yet characterized in cancer context. Given the importance of these switch genes, the objective of this project is to identify switch genes in Prostate adenocarcinoma gene expression data from the TCGA database. To reach the goal, using the SWIMmeR package, the data is filtered, the networks are found from that data and from those networks switch genes are found. As a final step, resilience analysis is executed. The ultimate acquired genes could be investigated for further research in Prostate adenocarcinoma .

2 Introduction

Nowadays, complex network analysis is becoming more and more important, gradually penetrating important aspects of biology and medicine. The disease is seldom a cause of mutation in one gene: it is rather a combination of the multiple interconnections between various cellular components(Nature Review Genetics). Thus, understanding the impacts of these relationships on the development of disease can result in discovery of new therapeutic targets. Research in cancer is not an exception in this case either. Cancer is frequently caused by variations in particular genes that make cells multiply unstoppably (Sanchez-Vega F et al., 2018). RNA sequencing is used more and more to understand the specific mechanism behind the process and ultimately to produce drugs that can be used to hinder specific cancer alterations. After tissue samples from normal and cancerous tissues are obtained, RNA sequencing is executed. In this project, statistical methods from the SWIMmeR package were applied on Prostate adenocarcinoma data from the TCGA database. The research in Prostate

adenocarcinoma is quite important since it is the most common form of cancer in men (Hartmann & Friess, 2017). The objective is to discover crucial cell nodes, namely switch genes that could further accelerate research in treatment of Prostate Carcinoma. Previous research will also be included in this paper.

3 Materials and Statistical Methods

3.1 Exploratory data analysis

The data could be acquired from the National Cancer Institute GDC Data Portal. For this task, as usual the SWIMmeR package was used. Data was formed into a dataframe and only subjects with both cancer and normal tissue gene expression were left as a result. Thus, out of the dataset with 550 samples only 104 samples were left (52 of both cancer and normal tissue). For each of the 52 samples, we have a matrix of count data $K \in \mathbb{N}_0^{I \times J}$, with superscripts K^N and K^C denoting normal tissue or cancer tissue data, respectively. Each column represents one of the J patients whose tissue was examined in the study and each row of the matrix indicates one of I genes considered in the study. Thus, each entry k_{ij} of a count matrix refers to the number of reads for gene i in the sample of a patient j.

Now that the number of samples is decreased, there are still 20531 genes left and these genes need to be preprocessed. Thus, genes with relatively small interquartile range are removed because the small IQR signifies that the values are less dispersed around the median. The IQR is the difference between the 75th and 25th percentiles of the data. To calculate the IQR, the data set is divided into quartiles, or four rank-ordered even parts via linear interpolation. These quartiles are denoted by $Q1$ (the lower quartile), $Q2$ (the median), and $Q3$ (the upper quartile). The lower quartile corresponds to the 25th percentile and the upper quartile corresponds to the 75th percentile, so IQR = $Q3$ – $Q1$ (Dekking et al.). In addition, the genes with a number of zeros greater than a certain threshold are discarded for a better analysis.

However, this is still not efficient since we still have to further filter our data. So, the next step is to find differentially expressed genes where the expression level between cancer and normal tissues differ significantly. To achieve this goal, firstly the logarithmic fold changes of the genes with respect to two tissue samples need to be found. We usually calculate a fold change of gene $i \in \{1, \dots I\}$ as the ratio of the average expression count of cancer tissue and normal tissue. The application of the $\log_2$ transformation to this measure yields the logarithmic fold change (LFC) of gene i defined as:

$$\mathrm{LFC}_i = \log_2 \left(\frac{\sum_{j=1}^{J} k_{ij}^C}{\sum_{j=1}^{J} k_{ij}^N} \right).$$

Fold changes with big absolute values imply that the expression level of the genes in the tissues differ with respect to a positive fold change indicating an up-regulated gene which means that this gene is expressed in cancer more than average. Thereby, LFC values can be used to find down-regulated or up-regulated genes.

For each gene i, we test the hypothesis:

$$H_0 : LFC_i = 0 \text{ vs. } H_1 : LFC_i \neq 0$$

using a Student's t-test. This means that under the null hypothesis the expression level of gene i is not affected by cancer. For each gene, p-value is calculated. Since usually a large number of genes is considered, the p values are adjusted using Benjamini and Hochberg method (Benjamini & Hochberg, 1995) which is used to avoid too many significant p-values due to multiple testing. All genes with adjusted p values smaller than a previously determined threshold can be regardes significantly differentially expressed genes.

3.2 Network Analysis

As a next step, co-expression networks are applied to the data of each tissue separately. To make a co-expression network, the similarity between the expression profile of a gene has to be quantified for each tissue sample. To measure the similarity of each pair of gene expression profiles, the absolute Pearson correlation coefficient can be applied (Niu et al., 2019):

$$s_{ih} = \left| \frac{\sum_{j=1}^{J}(x_{ij} - \bar{x_i})(x_{hj} - \bar{x_h})}{\sqrt{\sum_{j=1}^{J}(x_{ij} - \bar{x_i})^2 \sum_{j=1}^{J}(x_{hj} - \bar{x_h})^2}} \right|$$

with $\bar{x_i}$ describing the mean expression count of the gene i . This yields a similarity matrix $S = [s_{ij}] \in [0, 1]^{I \times I}$. s_{ij} being close to one means that the quantity of expression of gene i and j has high positive or negative correlation.

Now that we have a correlation matrix, we have to determine if we think the connection between two genes is strong. For that the signum-adjacency function is applied. This function is utilized to turn S with $s_{i,j} \in [0, 1]$ into an adjacency matrix $A = [a_{i,j}] \in \mathbb{R}^{I \times I}$ with. Another name for this technique for this procedure is also called hard thresholding and the a_{ij} are calculated by applying:

$$a_{ij} = \text{signum}(s_{ij}, \tau) = \begin{cases} 1, & \text{if } s_{ij} \geq \tau \\ 0, & \text{if } s_{ij} < \tau \end{cases}, i, j \in \{1, \dots J\}.$$

Here, $\tau \in (0, 1)$ is a threshold that we have to choose depending on the properties that we want to have (Niu et al., 2019). Generally the main diagonal of the adjacency matrix consisting of a_{ii} is set to zero. Regarding the main diagonal where all values will be 1, it is generally set to 0. Thus, this adjacency matrix can be transformed into a network graph. Each node in the graph represents a gene. There is a connection between node i and j, if the according entry a_{ij} in the adjacency matrix equals one. To find how connected node $i \in \{1, \ldots, I\}$ is to the other nodes of the same network, we can calculate the node's degree index based on the adjacency matrix. It is given by $\sum_{j=i}^{I} a_{ij}$ and thus gives the absolute number of connections of node i to other nodes of the graph.

Finally, the k-means algorithm is used to find communities in the correlation network. Given a set of observations $(x1, x2, ..., xn)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $S = S1, S2, ..., Sk$ so as to minimize the within-cluster sum of squares. Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where Ui is the mean of points in Si. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2$$

The equivalence can be deduced from identity $\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \sum (\mathbf{x} - \boldsymbol{\mu}_i)^T (\boldsymbol{\mu}_i - \mathbf{y})$ Since the total variance is constant, this is equivalent to maximizing the sum of squared deviations between points in different clusters (between-cluster sum of squares, BCSS), which follows from the law of total variance (Kriegel et al., 2016). Thereby, using this algorithm, communities of genes can be found.

3.3 Switch Mining

Once the communities of genes were found, the Average Pearson Correlation Coefficient (APCC) is assigned to each node of the correlation network. The APCC in this context is just a measure of coexpression of genes with nearest neighbors (Feiglin et al., 2014). For example, APCC for a gene with four neighbors will be as below:

$$APCC_i = \frac{p(i,j) + p(i,k) + p(i,l) + p(i,m)}{4}$$

As a result, we get three different types of hubs: date, party and fight-club. While party hubs interact with most of their partners simultaneously and date hubs bind their different partners at different times or locations, fight-club hubs are characterized by a marked negative correlation with the expression profiles of neighboring genes in the network (Palumbo et al., 2014).

Another interesting metric for gene analysis is the Clusterphobic coefficient (K_π) which measures the fear of being confined in the cluster. In other words, it evaluates the ratio of internal to external connections of a node and thus represents a measure of global connectivity. In general, A high value of denotes nodes having much more external than internal links (Feiglin et al., 2014):

$$K_\pi = 1 - (\frac{K_i^{in}}{K_i})^2$$

Also, we also have to find within module degre z_g which measures how much a gene is a hub within its own cluster:

$$z_g^i = \frac{k_i^{in} - \bar{k}_{C_i}}{\sigma_{C_i}}$$

where K_i^{in} is the number of links of node i to nodes in its module C_i . $\overline{K_{C_i}}$ and σ_{C_i} are the average and standard deviation of the total degree distribution of the nodes in the module C_i.

As a result, the within module degree and clusterphobic coefficient helps us to identify the switch genes. More specifically, if a gene is not a hub within its own cluster ($z_g < 2.5$), has many clusters outside its own neighbourhood ($K_\pi > 0.8$) and has a negative average weight of its incident links (APCC < 0), we consider it as a switch gene (Feiglin et al., 2014).

3.4 Resilience Analysis

In this step we check the response of the network to the removal of nodes or links. Network robustness is performed by studying the effect on the network connectivity of removing different types of nodes by decreasing degree. The total number of nodes to be removed must be equal to the total number of switch genes and the cumulative node deletion is carried out by type (i.e., total hubs, party hubs, date hubs, fight-club hubs, switch genes, and randomly chosen nodes). This step has to help us to understand the behaviour of switch genes with respect to arbitrary genes.

4 Results

4.1 Exploratory data analysis

With the help of the SWIMmeR we found the expression levels of the given genes. Firstly, the genes with a number of zeros greater than 75 % were discarded. Then, the threshold for interquartile range was selected at 0.11. To filter out the most expressed genes we used log2 fold change of 3.4 adjusted p-value of 0.05 as a threshold to fight with the multiple testing problem. The result of filtering is that we arrived at 417 differentially expressed genes which are depicted in the volcano plot in figure 1.

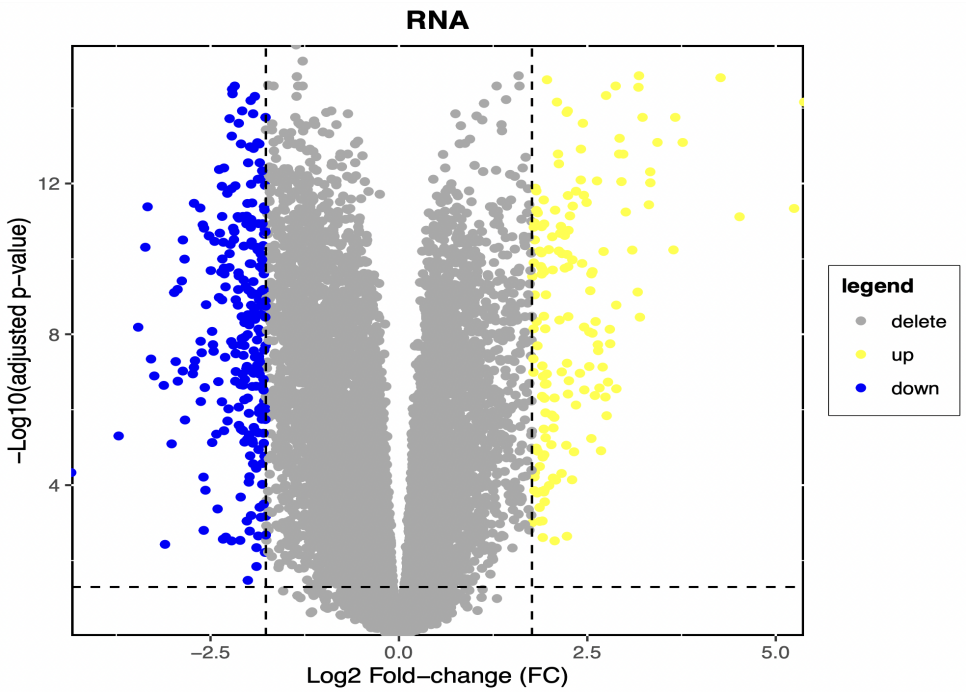

Figure 1: Volcano plot of the differentially expressed genes

4.2 Network Analysis

In the following part of the report, we used only the 417 differentially expressed genes previously identified for the calculation of (differ-ential) co-expression networks. We calculated the absolute Pearson correlation for each pair of genes and we got the similarity matrix as a result. In order to select an appropriate threshold, an adjacency matrix was calculated using hard thresholding for different threshold-candidates ranging from 0.5 to 0.9. For each potential threshold the degree index for each of the 417 genes was calculated. While the histograms for the degree distributions in the cancer tissue indicate a scale free network for all potential

thresholds, the histograms for the normal tissue do not. For this reason and after checking network integrity, we decided to use tau = 0.56 in the study (Figure 2) , even though the average degree index for the cancer tissue co-expression network decreases.

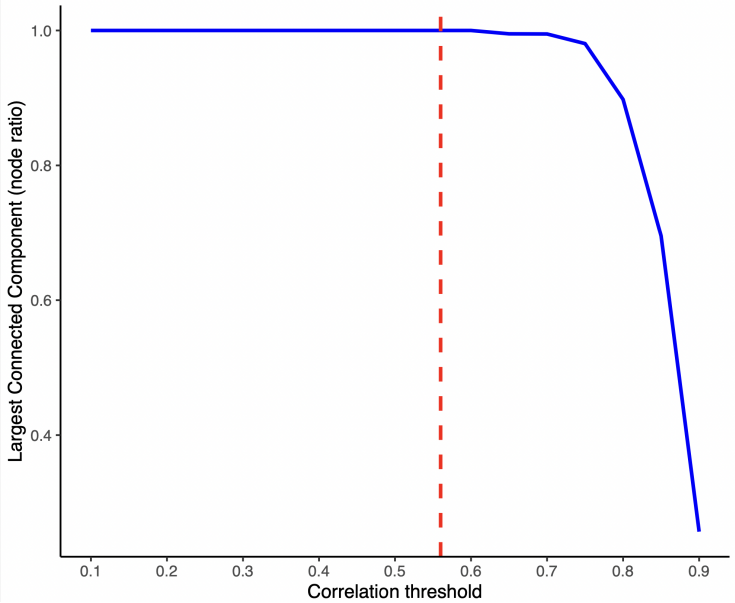

Figure 2: Network Integrity Plot

In figure 3, the histogram with the final threshold is chosen. Red bars are the selected highly correlated pairs and grey bars are the deleted poorly correlated pairs.

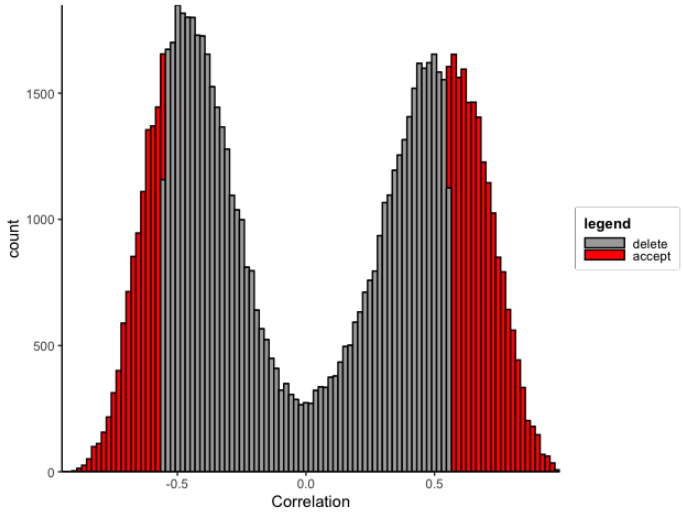

Figure 3: the Pearson correlation coefficient between the expression profiles of all pairs of genes

After adjusting the matrix of correlations to contain only highly correlated pairs, we apply the k-means algorithm to find communities of genes. We decided to set the number of clusters

to seven since it gave the least error for the maximum number of iterations allowed 100 and number of repetitions of clustering equal to 5.

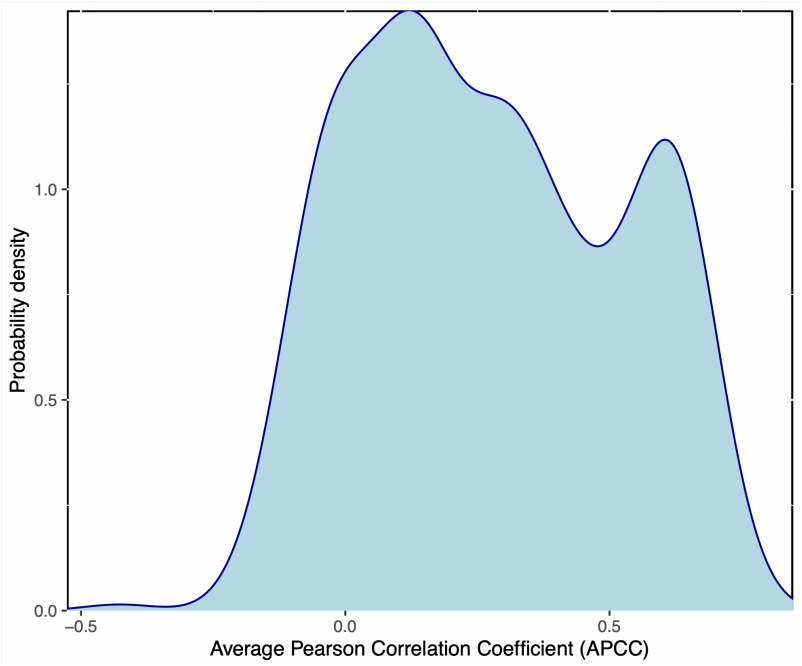

Figure 4: APCC distribution

4.3 Switch Mining

The curve represents the estimated probability density using a smoothing algorithm with a Gaussian kernel of the APCC values' distribution. Here it can be seen that there are many genes in party hubs but mostly genes are in date hubs and there are almost no genes in fight hubs.

After finding the within module degree and clusterphobic coefficient and using APCC values we can construct a heat cartography map as given in figure 5. In that figure, according to our metrics in the methods parts, switch genes are located in region 4, down the right corner.

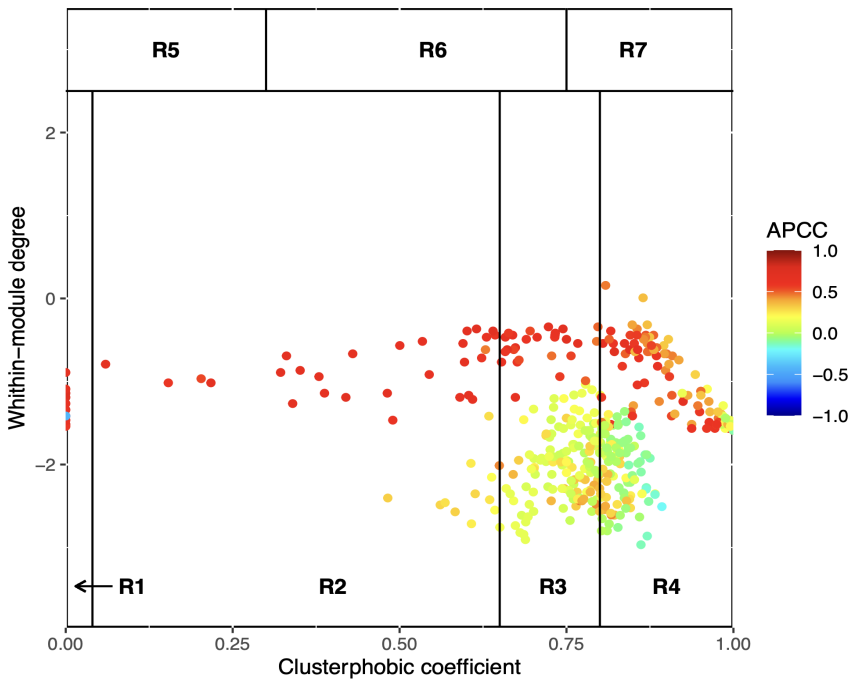

Figure 5: Heat cartography map

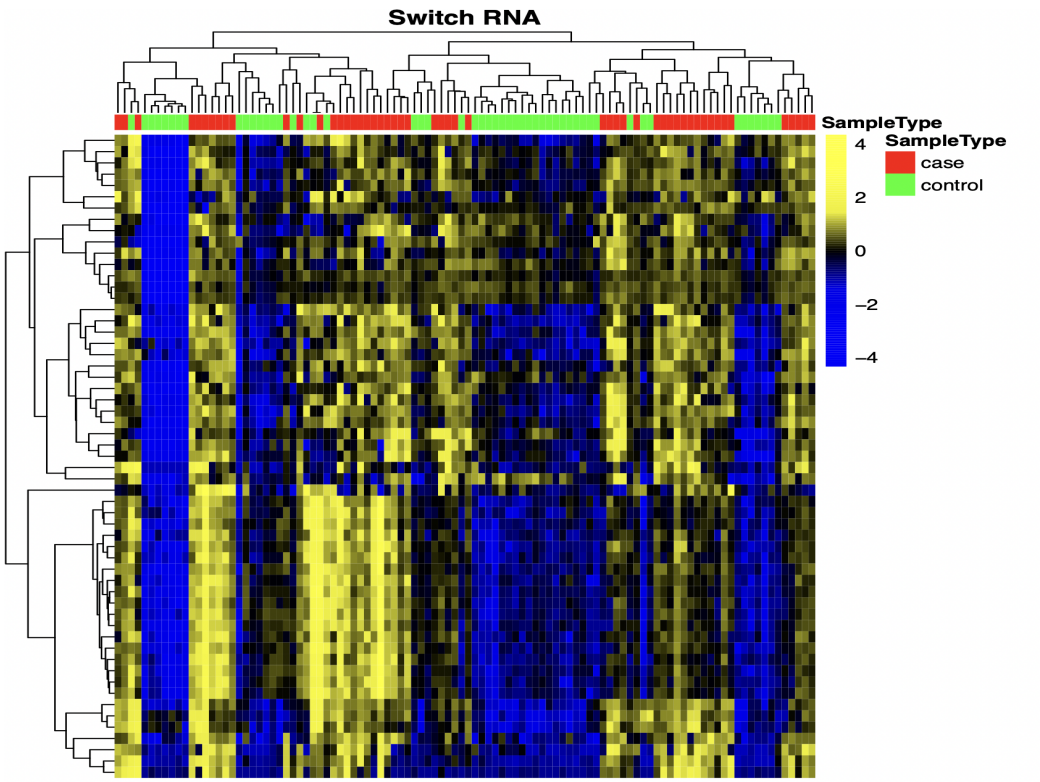

Figure 6: Heatmap of switch genes (RNAs)

In figure 6, rows refer to switch RNAs, columns refer to samples, and colors represent different expression levels that increase from blue to yellow. Thus, in contrast to other RNA genes on the heatmap in figure 8 in the appendix, in switch genes the discrepancy can be clearly seen with case samples being much more expressed. Thus, we found our switch genes.

4.4 Resilience Analysis

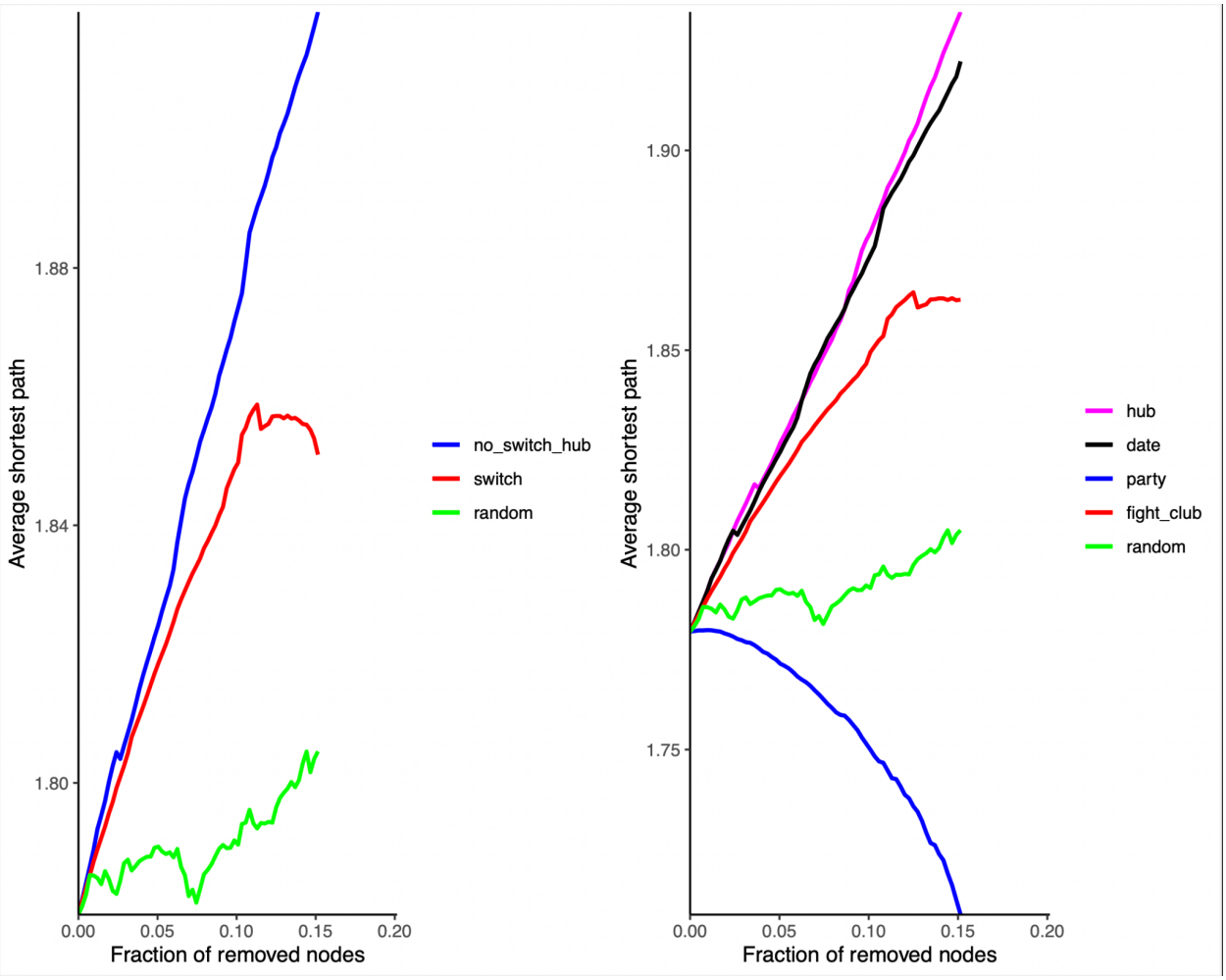

Figure 7: Network Resilience Plot

In figure 7, by average shortest path the average of minimum number of edges connecting them is implied. Thus, in the plot it is depicted how the average shortest path changes as the fraction of removed nodes increases. Immediately it can be seen that the switch genes behave completely differently from regular random genes: for random genes fraction of removed nodes just gradually increases average shortest path while for the switch genes the increase is almost exponential. What is also interesting is that for fight club hubs the average shortest path decreases which is most likely due to their low number.

5 Discussion

Clear differences in gene expression and co-expression between cancer and normal tissues were demonstrated. As a result, a total of 499 differentially expressed genes were identified, with some being up- and others being down-regulated in cancer tissue. Among these genes, we could identify 63 cancer tissue-hub genes. Since these genes might be connected to

the cause of uncontrollable cell division cycles in cancer cells, further investigation into these genes might be interesting.

To check if the chosen switch genes make sense, among all the switch genes we chose two with highest fold changes, namely ZIC2 (Zic Family Member 2) with a log fold change of 4.27 and HOXC6 with a log fold change of 3.43. Indeed, diseases associated with both genes include Lymphoma, Non-Hodgkin, Familial and Holoprosencephaly (Genecards.org). More interestingly, there are researches backing up connection of both genes to the development of prostate cancer. Homeobox (HOX) genes, which are involved in organ development and homeostasis, have been shown to be involved in normal prostate- and PCa development (Hamid et al., 2015). While ZIC2 and its overexpression has been found to be related to invasiveness and metastasis in nasopharyngeal, breast, prostate cancer, acute myeloid leukemia (Lv et al., 2021). Thus, it will be interesting to take a look at other genes from our list of switch genes and check the relationship between them and cancer types. And even these two genes could be studied from different perspectives to get a better idea about them.

Appendix:

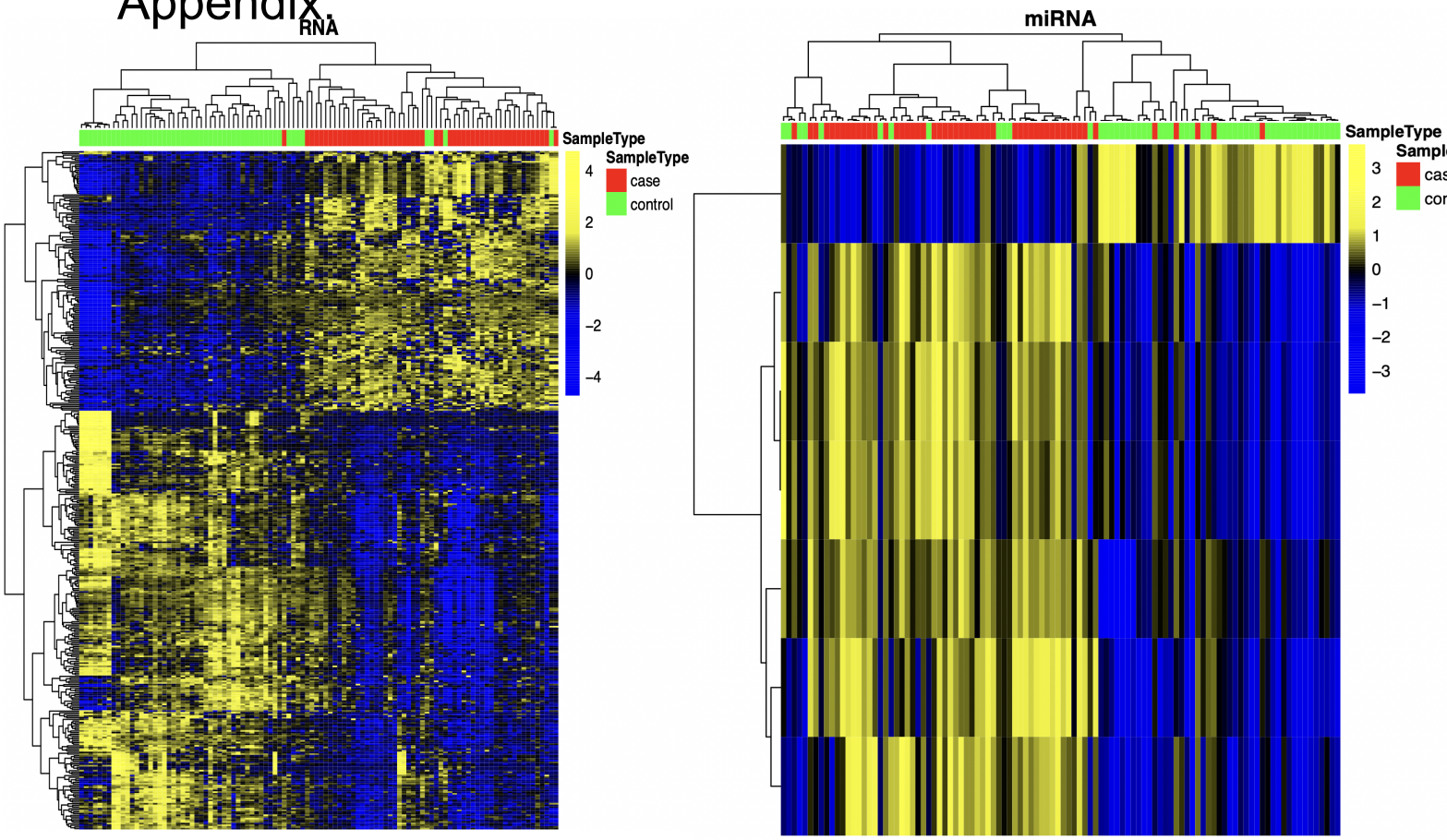

Figure 8: Heatmap of RNA and miRNA

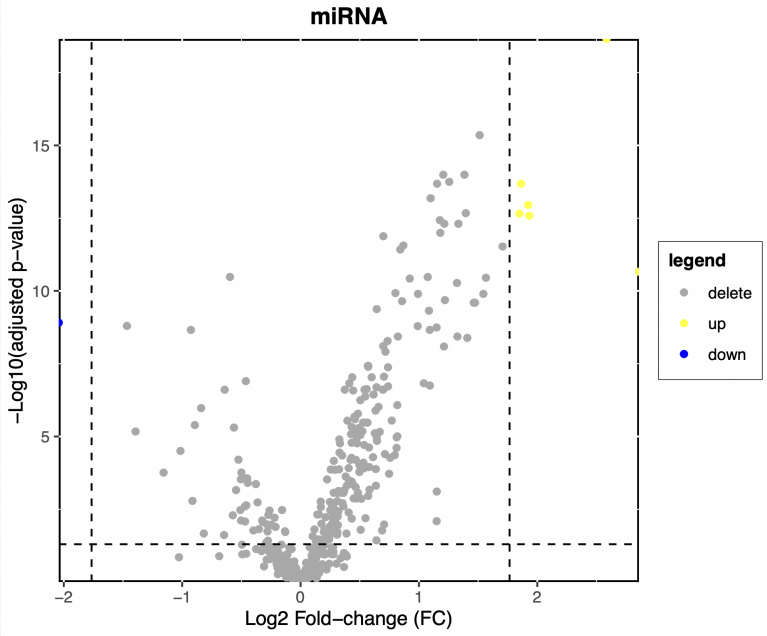

Figure 9: Volcano Plot miRNA

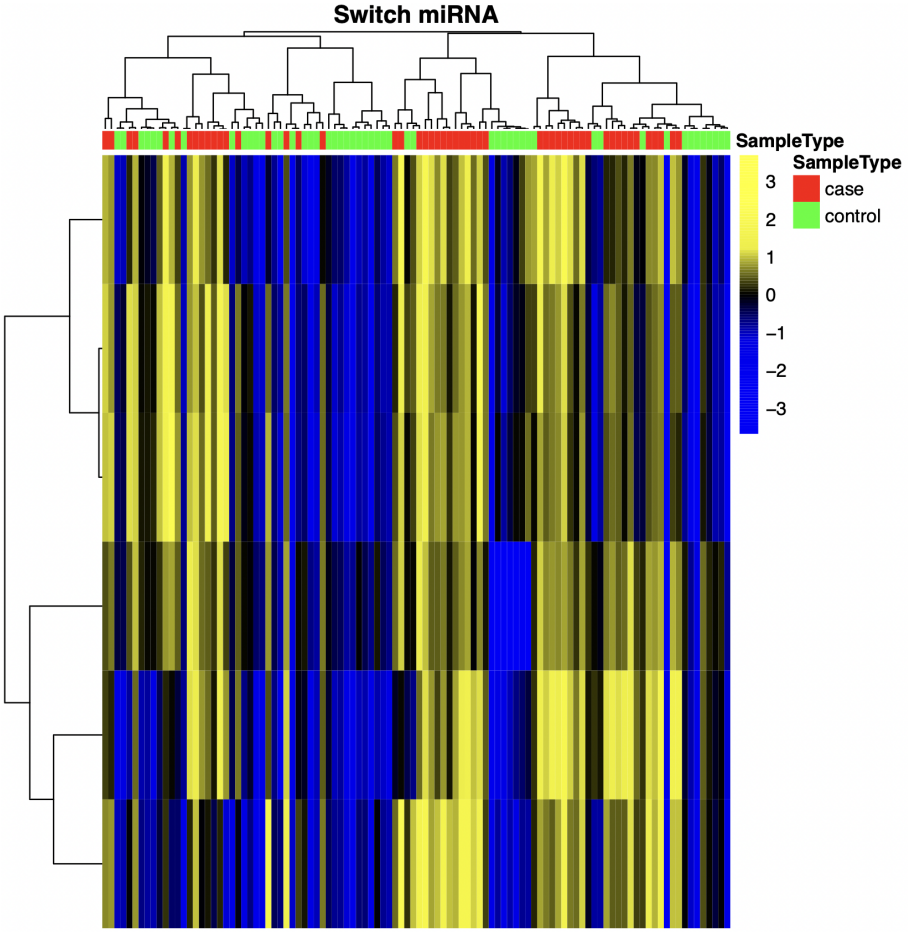

Figure 10: Heatmap of Switch RNA

Works cited:

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (n.d.). *A modern introduction to probability and statistics: Understanding why and how*.

Feiglin, A., Ashkenazi, S., Schlessinger, A., Rost, B., & Ofran, Y. (2014). Co-expression and co-localization of hub proteins and their partners are encoded in protein sequence. *Molecular BioSystems*, *10*(4), 787. https://doi.org/10.1039/c3mb70411d

Genecards.org. (n.d.). Retrieved December 23, 2021, from https://www.genecards.org/

Hamid, A. R., Hoogland, A. M., Smit, F., Jannink, S., van Rijt-van de Westerlo, C., Jansen, C. F., van Leenders, G. J., Verhaegh, G. W., & Schalken, J. A. (2015). The role of Hoxc6 in Prostate Cancer Development. *The Prostate*, *75*(16), 1868–1876. https://doi.org/10.1002/pros.23065

Hartmann, & Friess. (2017). *Prostate adenocarcinoma*. Prostate Adenocarcinoma - an overview | ScienceDirect Topics. Retrieved December 19, 2021, from https://www.sciencedirect.com/topics/medicine-and-dentistry/prostate-adenocarcinoma#:~:text=Prostatic%20adenocarcinoma%20(PAC)%20is%20the,as%20a%20precursor%20for%20PAC.

Kriegel, H.-P., Schubert, E., & Zimek, A. (2016). The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*, *52*(2), 341–378. https://doi.org/10.1007/s10115-016-1004-2

Lv, Z., Qi, L., Hu, X., Mo, M., Jiang, H., Fan, B., & Li, Y. (2021). Zic family member 2 (ZIC2): A potential diagnostic and prognostic biomarker for Pan-Cancer. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.631067

Nature Reviews Genetics. (n.d.). https://doi.org/10.1038/41576.1471-0064

Niu, X., Zhang, J., Zhang, L., Hou, Y., Pu, S., Chu, A., Bai, M., & Zhang, Z. (2019). Weighted gene co-expression network analysis identifies critical genes in the development of heart failure after acute myocardial infarction. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01214

Paci, P., Colombo, T., Fiscon, G., Gurtner, A., Pavesi, G., & Farina, L. (2017, March 20). *Swim: A computational tool to unveiling crucial nodes in complex biological networks*. Nature News. Retrieved December 19, 2021, from https://www.nature.com/articles/srep44797

Palumbo, M. C., Zenoni, S., Fasoli, M., Massonnet, M., Farina, L., Castiglione, F., Pezzotti, M., & Paci, P. (2014). Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch genes that induce major transcriptome reprogramming during grapevine development . *The Plant Cell*, *26*(12), 4617–4635. https://doi.org/10.1105/tpc.114.133710

Sanchez-Vega F;Mina M;Armenia J;Chatila WK;Luna A;La KC;Dimitriadoy S;Liu DL;Kantheti HS;Saghafinia S;Chakravarty D;Daian F;Gao Q;Bailey MH;Liang WW;Foltz SM;Shmulevich I;Ding L;Heins Z;Ochoa A;Gross B;Gao J;Zhang H;Kundra R;Kandoth C;Bahceci I;Dervishi L. (n.d.). *Oncogenic signaling pathways in the cancer genome atlas*. Cell. Retrieved December 19, 2021, from https://pubmed.ncbi.nlm.nih.gov/29625050/