Analysis of gene expression data for Uterine Corpus Endometrial Carcinoma

Dusica STEPIC, Omirbanu NURASSILOVA and Melis KAYMAZ[1]

Abstract

In this project, the aim was to identify Differentially Expressed Genes (DEGs) from the data collected in the two samples, the samples with Solid Tissue Normal and the samples with Primary solid Tumor. We performed several analyses, including building a binary adjacency matrix, based on Pearson's correlation coefficient in order to create a network between the genes in both conditions and find their hubs. Moreover, we also detected and identified some genes in the cancer hub that may be related to Uterine Corpus Endometrial Carcinoma (UCEC). We used the Python programming language[2] for the analysis and R for the data extraction.
Keywords: gene expression, transcriptome profiling, endometrial carcinoma

Introduction

For the analysis of the gene expression data, identifying differentially expressed genes (DEGs) based on p-value and fold change that is specified according to the disease we examine was performed firstly. Then, the gene co-expression networks with respect to cancer and control groups are calculated. This part was done by computing Pearson's correlation and constructing a binary adjacency matrix. As a third step, the degree index was calculated, and the top 5 percent nodes with the highest degree values (hubs) were found. The main purpose was to compare two hubs sets; cancer and normal tissue in order to characterize and gather more information about hubs related to cancer tissue. In addition to our conclusion, we examined the literature to determine whether there are any studies regarding the genes that we found in connection with UCEC.

Results and discussion

Data extraction

Data was downloaded from https://portal.gdc.cancer.gov/ using the R programming language and the TCGAbiolinks[3] package, an R/Bioconductor package for integrative analysis with GDC (Genomic Data Commons) data. The parameters needed to be specified are *data category*: Transcriptome Profiling; *data type*: Gene Expression Quantification; *workflow type*: HTSeq – FPKM and *Project ID*: TCGA-UCEC.
The following steps of the project were performed using Jupyter Notebook and the python programming language. After the gene expression data was imported, it was filtered by selecting just the patients for whom cancer and normal tissue files are available.

---

[1] University of Rome, Sapienza, MSc Data Science, Digital Epidemiology class 2019/2020, Group No: 1
[2] https://www.python.org
[3] https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html

Identifying Differentially Expressed Genes (DEGs)

The goal was to identify the Differentially Expressed Genes (DEGs) and reduce the number of genes from 56512 to a smaller subset of hundreds of genes. In order to decide the thresholds for Fold Value, we tested different configurations, by specifying different thresholds settings, to determine the best values to obtain a smaller subset with around 226 genes. Figures 3 and 4 show the Volcano plots with two different Fold Change values. This plot shows how the number of Fold Change we choose is decreasing the subsets of genes. By trial and error, we decided the appropriate number for the threshold.

Firstly, in order to identify DEGs and determine a smaller subset of DEGs, some calculations were performed along with the specified threshold setting for the measures. The first step was to calculate the Fold Change. Furthermore, we removed the genes that were 0 in both conditions for all the patients.

$$\text{Fold change} = \mathbf{log_2} \frac{\boldsymbol{Expr\ Cond}}{\boldsymbol{Expr\ Cond\ 2}}$$

In order to determine which genes are differentially expressed, first we took the proportion of cancer tissue and normal tissue for each gene, then calculated the log of 2. Now, every gene has a fold change number. As a threshold, 1.2 was not enough to reach a subset of few hundreds of genes, thus 4.7 was decided as the most appropriate threshold for our problem. This limitation gives us exactly 724 unique genes as the sample for performing t-tests among all those pairs of genes. In this part, only genes that are greater than the threshold are used. Since the two groups' mean are compared, whether they are equal or not for each gene type, we decided to use the t-test. Before that, we have to check if their variances are equal or not. If they are not, Welch's t-test, which is also known as unequal variances t-test, should be applied. In our case, a group with cancer tissue and normal tissue have different variances, so Welch's t-test[4] was conducted. As the p-value, 0.01 was chosen, and the gene pairs with p-value smaller than 0.01 were eliminated. Hence, we get 284 DEG pairs.

When multiple comparisons are needed, the correction method of p-values can be applied. In that case, False Discovery Rate (FDR) can be used as an expected proportion of type 1 errors, and it checks if the k ordered p-value is larger than $\frac{k \times a}{N}$ or not. FDR approach takes control for a low rate of false positives (FP), contrary to covering against making any FP result at all. For medical research, FDR means the time that the test results is "positive", but actually, that patient doesn't have the disease. After FDR correction was applied, 226 genes were obtained.

We sampled 3 genes from the part of the dataset labeled as DEG and the part of the dataset that is not DEG and plotted them to assure if they are different in two conditions. Figures 1 and 2 clearly show the difference between DEGs in two conditions.

Gene co-expression

Gene co-expression network is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them. First, we need to calculate Pearson's correlation between each pair of genes and build a correlation matrix. We may call this matrix as Similarity (co-expression) Score. Next, the significance threshold should be chosen

[4] https://en.wikipedia.org/wiki/Welch%27s_t-test

in order to create a network adjacency matrix. This matrix should be binary where a_{ij}(element of the matrix) is zero when the Pearson correlation coefficient (R) is less than the significance threshold, otherwise, a_{ij} should be one. In our case, R is specified as 0.7. In other words when the following condition $|a_{ij}| > 0.7$ is satisfied then an element in the adjacency matrix is 1, otherwise it is 0.

```
Name: Gene co-expression network from adjacency matrix
Type: Graph
Number of nodes: 226
Number of edges: 2754
Average degree:  24.3717
```

Image 1 - Summary of binary adjacency matrix for whole network

Image 1 above shows the number of nodes, edges, and the average degree of the reduced subset of DEGs from the network created from the adjacency matrix.

Degree index and hubs

In undirected graphs, the degree of node *i* (where $i \in 1, \ldots, N$ and N number of nodes) is the total number of connections with other vertices. It measures the importance of a node within the network. The greater the degree, the more important is the presence of that node for the whole system.

$k_i = \sum_{j=1, i \neq j}^{N} a_{ij}$, $k_i \in$ [0, N-1] where $k_i$represents the degree of node i

In the directed graph, the degree can be split into the in-degree and out-degree:

In-degree of node i (where $i \in 1, \ldots, N$ and N number of nodes) is the total amount of links incoming to the vertex i. Similarly, the out-degree of nodes i (where $i \in 1, \ldots, N$ and N number of nodes) is the total amount of links outgoing from the vertex i. And the degree of node *i* is the of in-degree and out-degree.

We can obtain the degree of the network by summing all k_i values which are the degrees of each node.

Since our graph doesn't have any direction between edges, we can say that it is an undirected graph. Moreover, when we compute in-degree and out-degree for both conditions, they are all equal. Thus, we also prove that our graphs are undirected. We found degrees 978 and 2320 for the cancer tissue and control group respectively.

After that, we found hubs (5% of the nodes with highest degree values), which means top 5 of the most connected nodes related to two conditions.

```
[('ENSG00000130176', 22),
 ('ENSG00000124212', 21),
 ('ENSG00000154553', 21),
 ('ENSG00000146477', 21),
 ('ENSG00000133392', 20),
 ('ENSG00000163431', 20),
 ('ENSG00000175084', 20),
 ('ENSG00000121871', 20),
 ('ENSG00000235782', 19),
 ('ENSG00000254959', 19)]
```

```
[('ENSG00000022267', 41),
 ('ENSG00000172403', 39),
 ('ENSG00000198523', 39),
 ('ENSG00000065534', 38),
 ('ENSG00000077157', 38),
 ('ENSG00000254510', 37),
 ('ENSG00000138944', 35),
 ('ENSG00000163431', 34),
 ('ENSG00000130176', 31),
 ('ENSG00000149596', 30)]
```

Image 2- 5% of nodes with the highest degree values for cancer group and control group respectively

Comparison of hubs and characterizing hubs of cancer tissue

To compare the hubs we can simply look at the difference between the list of genes. However, both of the hubs have 8 unique genes, and only 3 in common. We may conclude that discrimination between cancer tissue and normal tissue is made clear.

The following list of genes are found related to UCEC based on our analysis: **ENSG00000130176** (Calponin 1 - CNN1), **ENSG00000124212** (Prostaglandin I2 synthase - TGIS), **ENSG00000154553** (PDZ and LIM domain 3 - PDLIM3), **ENSG00000146477** (Solute carrier family 22 member - SLC22A3), **ENSG00000133392** (Myosin heavy chain 11 - SMHC,SMMHC), **ENSG00000163431** (Leiomodin 1 - LMOD1), **ENSG00000175084** (Desmin - DES), **ENSG00000121871** (SLIT and NTRK like family member 3 - SLITRK3), **ENSG00000235782** (AL031429.1) and **ENSG00000254959** (INMT-MINDY4 readthrough). The list is organized in relevance order i.e. CNN1 is the most relevant one to UCEC.

To show the consistency of our analysis, some genes are checked whether there is already research that shows a link between the UCEC and the gene in literature or not. Our top gene in hubs, *Calponin h1* may function as a tumor suppressor in leiomyosarcoma which is a type of soft tissue sarcoma. Clinically, transfer of calponin h1 complementary DNA into poorly differentiated leiomyosarcoma cells may be of potential therapeutic value through the induction of a normal, differentiated cellular phenotype.[5]

Furthermore, one of the most important gene in the hub of the network is *Prostacyclin synthase* and thromboxane synthase signaling via arachidonic acid metabolism affects a number of tumor cells survival pathways such as cell proliferation which is rapid reproduction of a cell, apoptosis which means a form of programmed cell death that occurs in multicellular organisms, tumor cell invasion, and metastasis, and angiogenesis.[6] Angiogenesis refers to blood vessel formation. Tumor angiogenesis is the growth of new blood vessels that tumors need to grow.

Conclusion

First of all, we identified DEGs with a 4.7 Fold Change and 0.01 p-value. That gave us 284 unique DEG pairs to use them in further steps. We applied a correction method; FDR and after that, we got 226 genes. Furthermore, we established a network between the genes by turning a binary outcome of Pearson's correlation result. In order to do that, R's threshold is specified as 0.7. We got a huge and complex network; our main interest here was 5% of the nodes with the highest degree values which are hubs. When we compare hubs for cancer and control groups, we realized that there is an intersection with only 3/11 genes between them, and 8 disjoint nodes. In order to complete an analysis, we researched on Internet to check that if there are paper(s) claiming that some of the genes in our cancer hub are found relevant to UCEC. According to Horiuchi et al. (1999), **Calponin h1**, which is the most relevant gene in our cancer hub, has a dominant role in human uterine leiomyosarcoma. Also, Cathcart MC et al. (2010) emphasized that the role of downstream effectors of **PGIS** that is our second important gene in our cancer hub activity in tumor growth and progression. All in all, we can conclude that there is a clear connection to the genes that we found and UCEC.

[5] https://www.ncbi.nlm.nih.gov/pubmed/10328110
[6] https://www.ncbi.nlm.nih.gov/pubmed/20122998

Tables and Figures

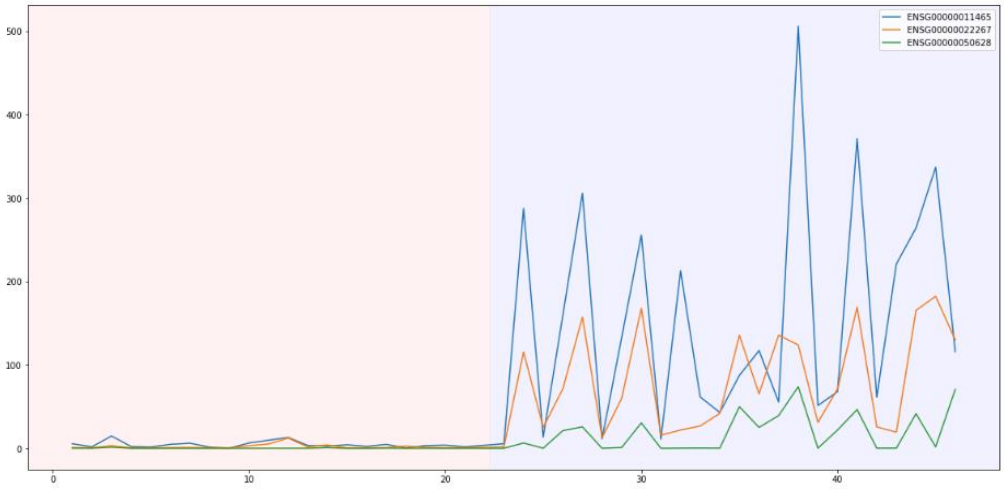

Figure 1 - Expression value of the sampled Differentially Expressed Genes (DEG) in two conditions

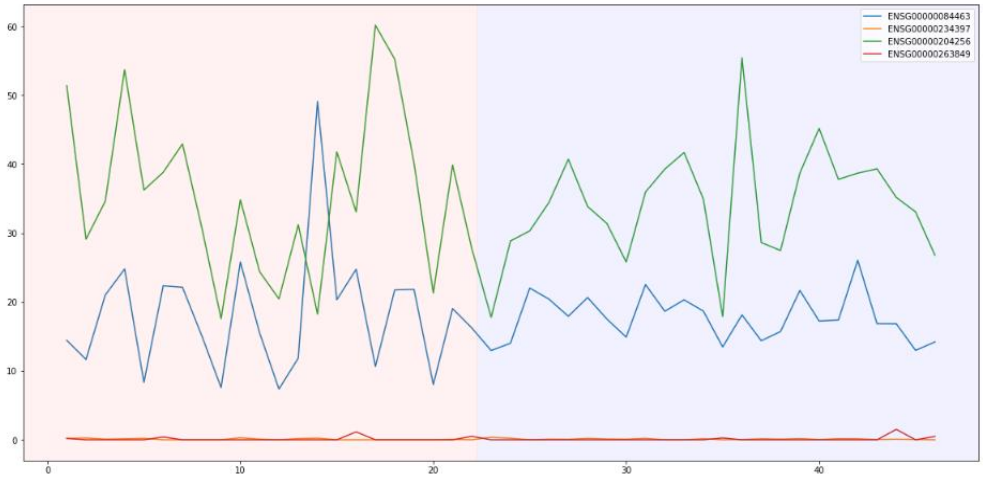

Figure 2 - Expression value of the sampled Non-Differentially Expressed Genes (DEG) in two conditions

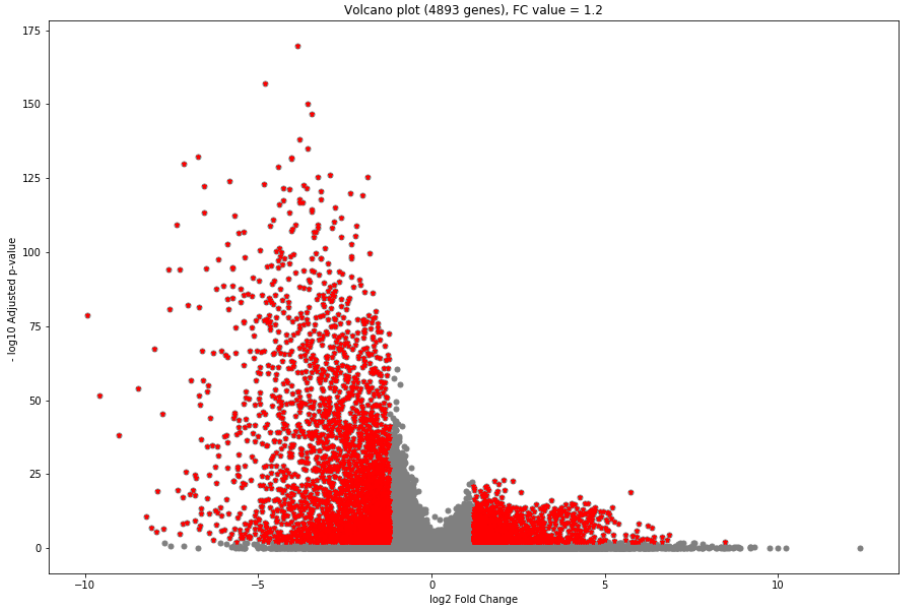

Figure 3 -Volcano plot, FC value = 1.2

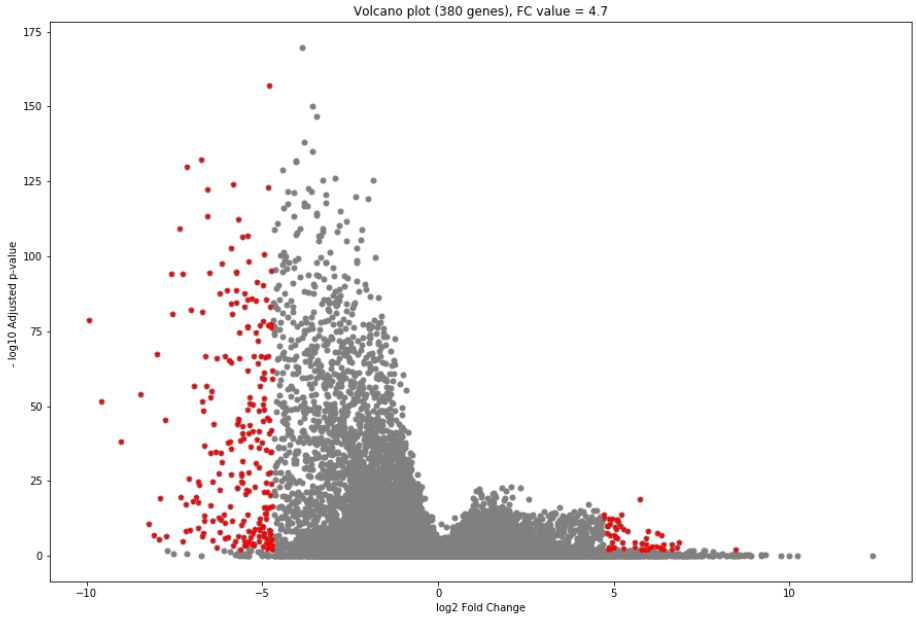

Figure 4 - Volcano plot, FC value = 4.7

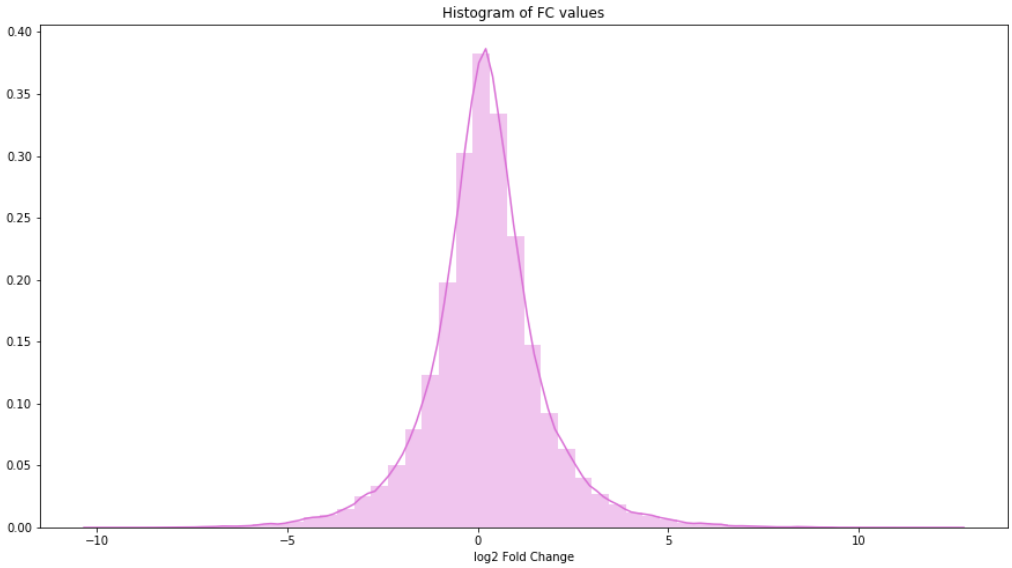

Figure 5 - Histogram of FC values

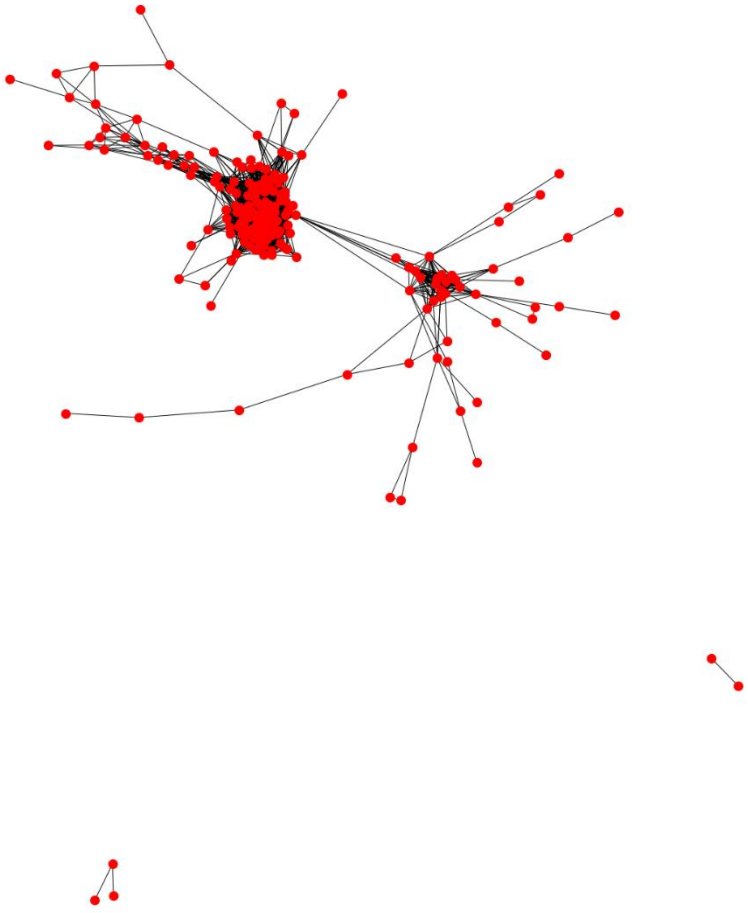

Figure 6 - Graph of the whole network (solid tissue normal and primary solid tumor)

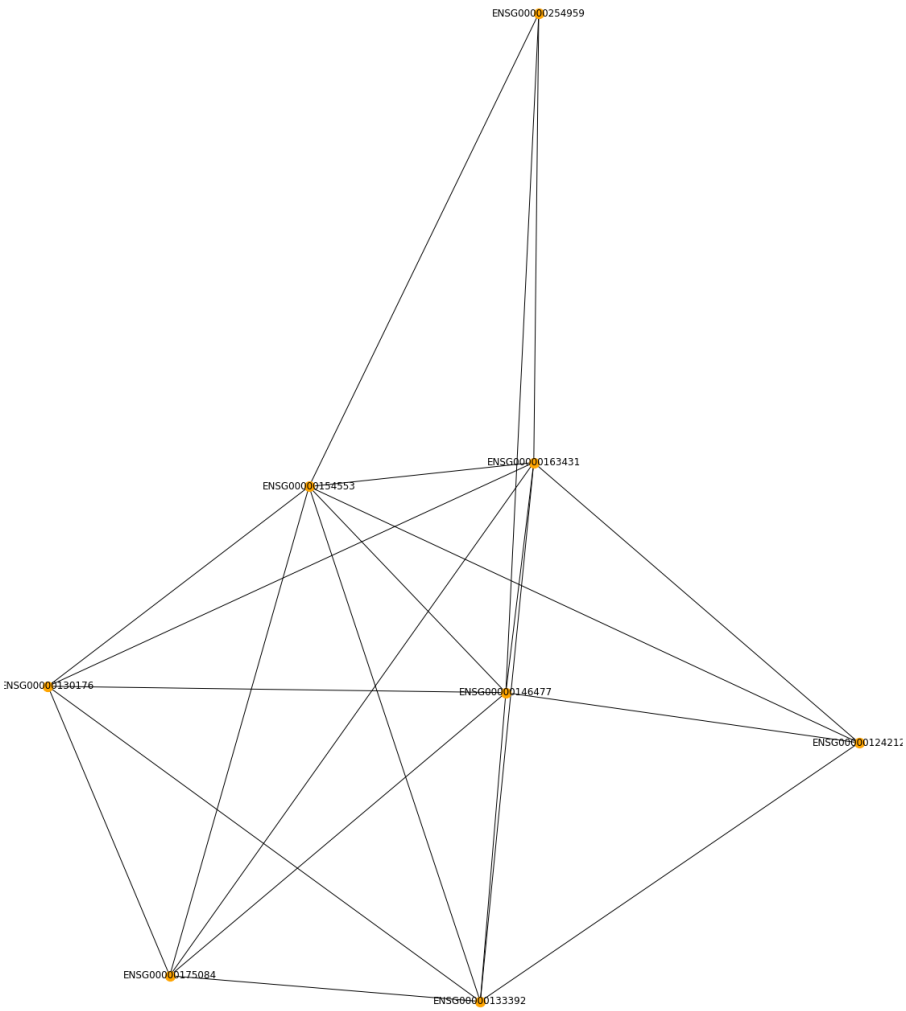

Figure 7 - Cancer hub (hub of the cancer tissue genes)

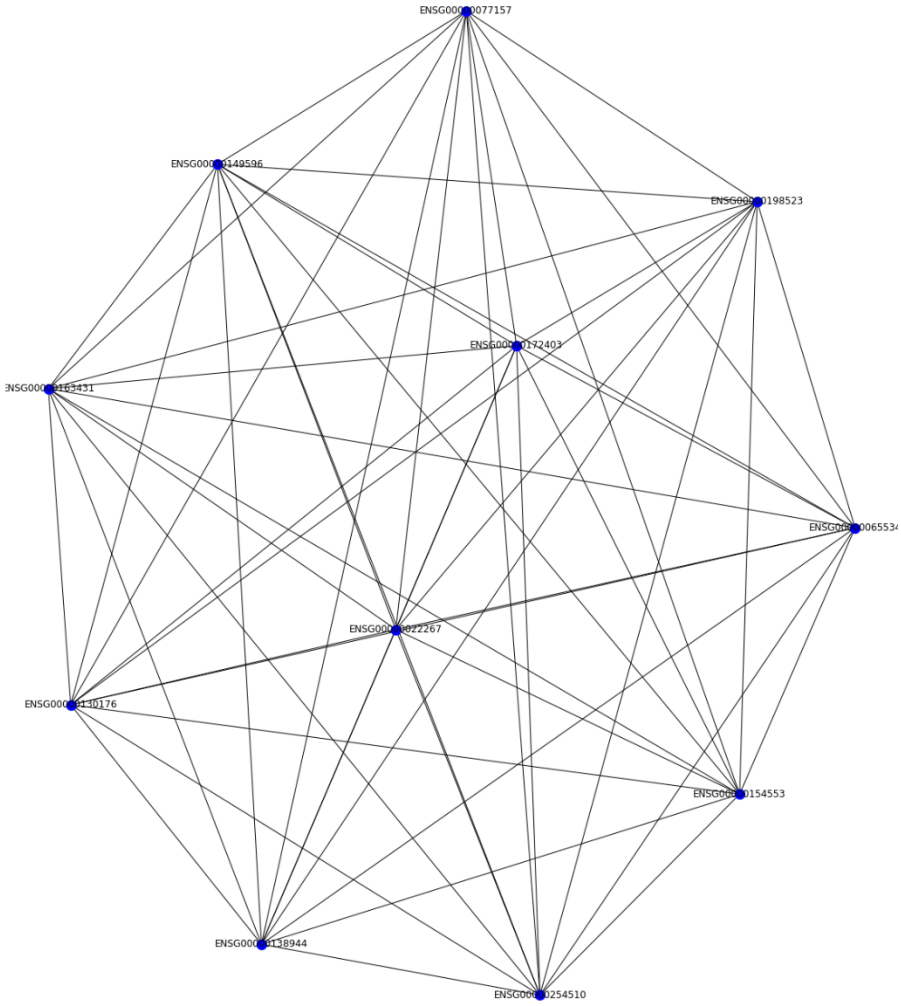

Figure 8 – Normal tissue hub