Differential Analyses of Gene Expression to Investigate Sexual Dimorphism in Liver Hepatocellular Carcinoma

Di Nino Leonardo 1919479

Di Poce Giuseppe 2072371

June 2024

Abstract

This project aims at conducting a study on sexual dimorphism in liver hepatocellular carcinoma leveraging some tools from bioinformatics. In particular, the purpose was to identify Differentially Expressed Genes between cancer condition and control ones in both male and female patients, trying to pick out patterns and structures arising between the two groups. Further tools from network science have been deployed to provide deeper investigations, paying special attention to co-expression network and patient similarity networks: the analysis of the degree distribution and the community structures on these networks unveils properties only encoded in pairwise relations between entities.

1 Introduction

Gene expression is a foundational process in biology, whose activation as final outcome yields a protein. In this sense, it represents a precious tool in investigating many problems in biology, since protein synthesis is related to almost any primitive function of a cell.

The aim of this project is to apply tools from statistics and network science to gene expression data collected in the context of TCGA project for a specific disease. We want to study gene expression for male and female populations with respect to liver hepatocellular carcinoma, trying to find rising patterns under cancer and normal conditions.

In particular, we are going to use basic statistics summaries to identify differentially expressed genes, namely those genes whose expression present statistically significant differences between two different experimental conditions.

Following, we are going to apply network science models and algorithms to investigate the underlying connectomy of the gene expression and the significance of their hidden pairwise relations through differential co- expressed networks. The same reasoning will be applied to patients accordingly to the correlation between the gene expressions of two distinct individuals, defining and analysing patient similarity networks.

The full pipeline, from the data collection to the visualization, has been performed in R.

2 Materials and methods

2.1 Data

The data of interest comes from the TCGA-LIHC repository on The Cancer Genome Atlas. In particular, we were interested in the gene expression quantification linked to both primary tumor and solid tissue normal samples inside the transcriptome profiling. In the following we'll refer to the two considered experimental conditions as C and N. Moreover, we used records related to both genes and patients to retrieve additional information: in particular, we split our data according to the gender of our considered samples, that we'll refert to as M and F.

In the initial preprocessing we normalized data leveraging DESeq2 pipeline using *median of ratios* method applied within each condition, than we retrieved data related only to some specific patients and genes.

2.2 Differentially Expressed Genes (DEGs)

In order to identify differentially expressed genes, we retrieved two distinct metrics.

The first metric we collected is fold change, which quantifies differences in expression for a gene G across two different experimental conditions C and N: in particular, in order to deal with the multiple observations, we pooled the patients through sample averaging, so being $e(G)_{C,i}$ the expression for patient i in condition C and $e(G)_{N,i}$ the expression for patient i in condition N, we set:

$$FC(G)_M = \log_2 \frac{\sum_{i=1}^{|M|} e(G)_{C,i}}{\sum_{i=1}^{|M|} e(G)_{N,i}}$$
$$FC(G)_F = \log_2 \frac{\sum_{j=1}^{|F|} e(G)_{C,j}}{\sum_{j=1}^{|F|} e(G)_{N,j}}$$

The second metric we collected is the p-value for t-test between, again, averages across patients between the two conditions for each gene: in this case we needed a correction for multiple observations, and we decided for false discovery rate correction.

Then, we set a threshold for each of the metrics: in particular, we considered a gene to be differentially expressed between conditions C and conditions N if $|FC(G)| \geq 1.2$ and $p(G) \leq 0.05$. Finally, once we collected the DEGs for the two populations of interest, we investigated their overlapping.

2.3 Differential co-expressed network

We focused on two task in this case:

- We built the differential co-expressed network comparing cancer and normal condition for each sex;
- We built the differential co-expressed network comparing male and female population and cancer condition.

The interest in applying network science models to gene expression is related to the network structure per se: if the graph shows a degree distribution comparable with a scale-free structure, hubs are highly likely to arise, and these hubs might be genes of interest for advanced analysis.

In order to build a differential co-expressed network, we start from a correlation matrix ρ retrieved within each condition of interest, namely A and B. Then we apply the Fisher Z-transform to these correlation matrices: applying the hyperbolic arcotangent to correlation measures solve the high skewness of this statistic, yielding a random variable that approximately behaves like a Gaussian random variable.

$$z_{i,j} = \frac{1}{2} \log \frac{1 + \rho_{i,j}}{1 - \rho_{i,j}} \tag{1}$$

Finally, we compute the Z score comparing the two conditions:

$$Z_{i,j} = \frac{z_{i,j}^A - z_{i,j}^B}{\sqrt{\frac{1}{n^A - 3}} + \sqrt{\frac{1}{n^B - 3}}} \tag{2}$$

Interestingly, the variance-stabilizing term are related only to the sample size of the two population of interest within the two conditions considered for the analysis.

In the end, we threshold the Z scores to retrieve an adjacency matrix: $A_{i,j} = \mathbb{1}(|Z_{i,j}| \geq \overline{Z})$.

2.4 Patient Similarity Network (PSN)

In order to compute the Patient Similarity Network (PSN) we need a measure of similarity between the gene expression. For this section of the analysis, we rebuild a dataframe with all our patients, dropping the distinction based on gender, and we computed the Pearson correlation coefficents and the respective p-values between gene expressions. After multiple comparison normalization adjustment through FDR, we realized that thresholding the p-values would have yielded a fully connected network, given their magnitude. So we decided to threshold the correlation matrix such that the adjacency matrix of the final output graph is $A_{i,j} = \mathbb{1}(|C_{i,j}| \geq \overline{C})$.

Subsequently, we applied Louvain community detection algorithm to this network. Louvain algorithm is a 2-phases bottom-up clustering algorithm: the first step is a modularity optimization over the existing communities, aggregating communities if this aggregation yields the best improvement in the modularity; the second is a graph rebuilding step based on supernodes corresponding to the retrieved communities: the procedure is repeated until no further improvements are possible.

Finally, once we retrieved our communities, we compare them with our clinical records, paying particular attention to the fitting with the gender division.

3 Results and discussion

3.1 Differentially Expressed Genes (DEGs)

From our analysis on differential expression:

- We retrieved 67 DEGs among the male population (31 are up-regulated, 36 are down regulated);
- We retrieved 61 DEGs among the female population (34 are up-regulated, 27 are down regulated).

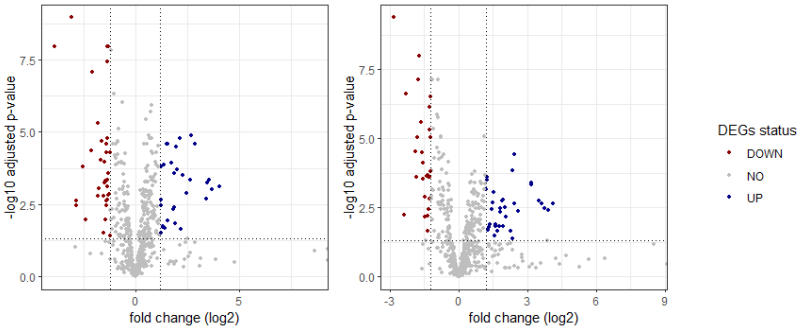

Figure 1: Volcano plot for DEGs with respect to fold change and p-value for t-statistics between cancer and normal tissue for male patients (on the left) and female patients (on the right)

ENSG00000004799.8	ENSG00000067082.15	ENSG00000073756.12	ENSG00000091879.14	ENSG00000095002.15	ENSG00000097046.13
ENSG00000101412.13	ENSG00000104635.15	ENSG00000106366.9	ENSG00000106462.11	ENSG00000106993.12	ENSG00000107566.14
ENSG00000107864.15	ENSG00000109805.10	ENSG00000113594.10	ENSG00000114346.14	ENSG00000116128.11	ENSG00000117399.14
ENSG00000117632.23	ENSG00000125257.16	ENSG00000130164.14	ENSG00000131747.15	ENSG00000134243.12	ENSG00000134294.14
ENSG00000136158.12	ENSG00000138160.7	ENSG00000138180.16	ENSG00000139318.8	ENSG00000140044.13	ENSG00000147889.18
ENSG00000150907.10	ENSG00000154065.17	ENSG00000155090.15	ENSG00000164045.12	ENSG00000164761.9	ENSG00000165757.9
ENSG00000171848.16	ENSG00000176597.12	ENSG00000177606.8	ENSG00000178999.13	ENSG00000185652.12	ENSG00000187498.16

Table 1: List of DEGs overlapping between men and women data

From our records on genes, we can move towards more advanced analysis. For example, among the 23 overlapping up-regulated genes many of them are located in chr10 and chr1, which could be linked to diseases of the endocrine system and in particular the liver, even in carcinomic forms: this could be the start for the investigation about the factors specifically related to the liver hepatocellular carcinoma.

3.2 Differential co-expressed network

3.2.1 Differential co-expressed network comparing cancer and normal condition

We realized that thresholding the Z-scores with a cutoff at 3 would have yielded a network with too many disconnected components: given the fact that the graph is already a small one, we preferred to retrieve a denser network, but also a more connected one setting the cutoff at $\overline{Z} = 2$. Taking into account distinctly male and female population we observe the following behaviour of degree distribution:

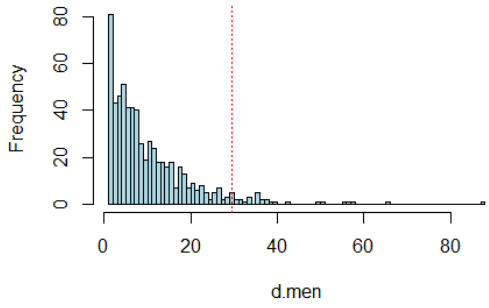

Figure 2: Male network degree distribution;

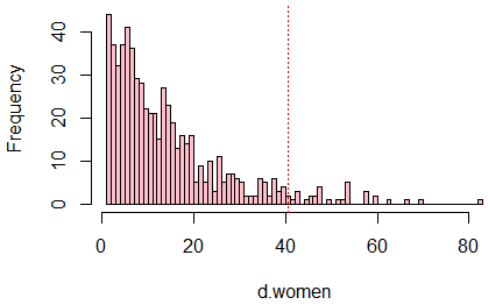

Figure 3: Female network degree distribution;

We want to assess if the degree distributions above show a *scale free* behaviour, characterized by highly uneven distribution of connections among its nodes [1]. In these networks some nodes, namely hubs, have many connections, while the vast majority of nodes have relatively few. We recall the power low distribution followed by scale free network nodes as $P(k) \sim Ck^{-\gamma}$: scale-free property is satisfied if $2 < \gamma < 3$, otherwise we end up in an *anomalous regime* if $\gamma < 2$ or we are back in a *random regime* if $\gamma > 3$.

In regime we could investigate the degree behaviour looking to the relationship between the number of nodes and their connections when both axes are plotted on a logarithmic scale. Infact, in this log-log transformation, the *power law* distribution transforms into a linear relationship as:

$$\log P(k) = \log C - \gamma \log k$$

so that we could fit a regression line to this transformed data to evaluate approximately the parameter of the distribution. However, the dimension of the network makes us question this approach, that would yield low statistically significance when interpreting the result. So we applied maximum likelihood estimation to our observed degrees and retrieved a value for γ equal to 4.39 for male degree distribution and 2.66 for female degree distribution: still its significance is to be assessed given the small data regime, but up to this we may conclude that, despite they are both well fitted by a linear model, only the female degree distribution shows a scale-free structure.

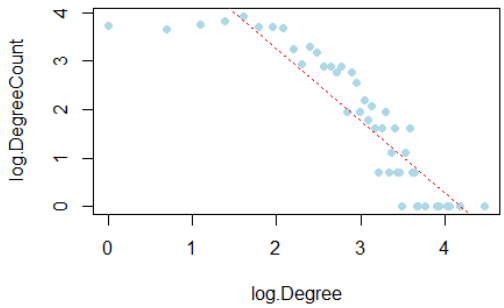

Figure 4: Male degree distribution in log-log scale

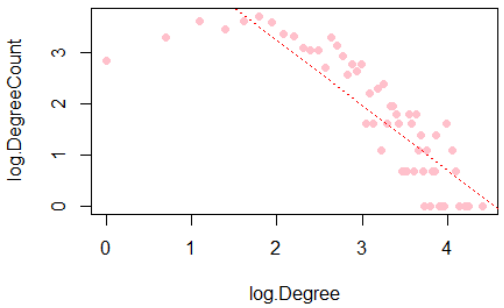

Figure 5: Female degree distribution in log-log scale

Additionally, we report the hubs we retrieved as the nodes whose degree was higher than the 95-th percentile of the distribution:

[1] Please notice that when considering log-log analysis we only considered the respective largest connected component of each network, since connectedness was required for the MLE for the power law distribution.

ENSG00000059758.8	ENSG00000069702.11	ENSG00000100644.17	ENSG00000101187.16	ENSG00000101413.12	ENSG00000105835.12
ENSG00000106993.12	ENSG00000108799.13	ENSG00000109320.13	ENSG00000113916.18	ENSG00000114861.23	ENSG00000115758.13
ENSG00000116984.14	ENSG00000117500.13	ENSG00000124216.4	ENSG00000126261.13	ENSG00000134294.14	ENSG00000136158.12
ENSG00000136448.13	ENSG00000137193.14	ENSG00000138685.17	ENSG00000138814.17	ENSG00000147649.10	ENSG00000152661.9
ENSG00000153147.6	ENSG00000155096.14	ENSG00000164961.16	ENSG00000165312.6	ENSG00000169398.19	ENSG00000173812.11
ENSG00000177606.8	ENSG00000184203.8	ENSG00000189376.12			

Table 2: List of hubs in male data

ENSG00000068878.15	ENSG00000076826.10	ENSG00000091136.15	ENSG00000095002.15	ENSG00000103222.20	ENSG00000107560.12
ENSG00000113594.10	ENSG00000113916.18	ENSG00000115392.12	ENSG00000116127.19	ENSG00000116128.11	ENSG00000120008.16
ENSG00000124496.12	ENSG00000136379.12	ENSG00000137193.14	ENSG00000138078.16	ENSG00000139793.18	ENSG00000140443.15
ENSG00000141582.15	ENSG00000144580.14	ENSG00000147251.15	ENSG00000156515.24	ENSG00000162711.18	ENSG00000164331.10
ENSG00000164961.16	ENSG00000169554.22	ENSG00000171492.14	ENSG00000171791.14	ENSG00000173801.17	ENSG00000180447.7
ENSG00000182481.10	ENSG00000185652.12	ENSG00000196628.20	ENSG00000198218.11		

Table 3: List of hubs in female data

We identified the following genes as overlapping hubs, that might be interesting for a joint analysis of the factors related to this disease since they show to be relevant in both male and female population:

- **BCL6** (ENSG00000113916.18), whose relevance in hepatocellular cancer is studied in [2]
- **PIM1** (ENSG00000137193.14), whose relevance in hepatocellular cancer is studied in [3]
- **WASHC5** (ENSG00000164961.16), low-specificity marker for liver cancer[4]

3.2.2 Differential co-expressed network comparing male and female population under cancer condition

This point in the analysis is very important, because it gives us another glimpse on differences in gene expression between the two genders when considering cancer tissue. In this section we retrieved a differential expression network for the cancer condition, using the gender division as the two hypotesis to retrieve the Z-scores. We mention that keeping $\overline{Z} = 2$ and so $A_{i,j} = \mathbf{1}(|Z_{i,j}| \geq \overline{Z})$, the following has evidence:

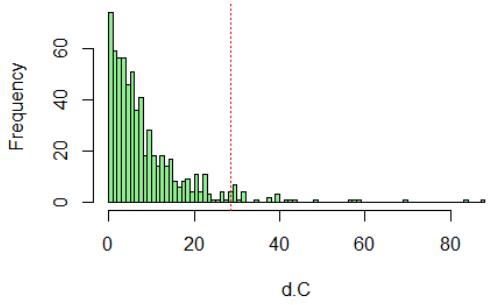

Figure 6: Cancer degree distribution;

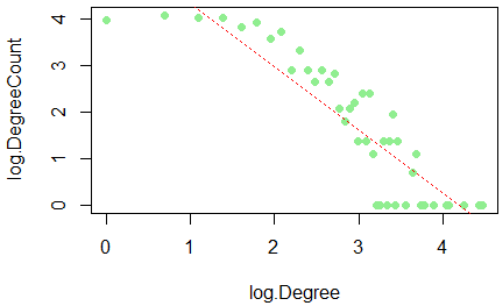

Figure 7: Cancer degree distribution in log-log scale;

Also this network has a value for the power-law MLE too high to state a scale-free behaviour: infact, we got $\hat{\gamma} = 3.95$. These are the hubs we retrieved:

These hubs should yield markers whose expression is significantly different between male and female patients. Interestingly, some of these hubs were already detected in previous analysis, while there are no hubs fully overlapping over this network and the two previous ones: this might asses the sexual dimorphism of the disease, since we retrieved different relevant genes using the two genders as differential conditions under cancer tissue hypotesis. In particular for the male population we have:

[2]Wang W, Huang P, Wu P, Kong R, Xu J, Zhang L, Yang Q, Xie Q, Zhang L, Zhou X, Chen L, Xie H, Zhou L, Zheng S. BCL6B expression in hepatocellular carcinoma and its efficacy in the inhibition of liver damage and fibrogenesis. Oncotarget. 2015 Aug 21;6(24):20252-65.

[3]Leung CO, Wong CC, Fan DN, Kai AK, Tung EK, Xu IM, Ng IO, Lo RC. PIM1 regulates glycolysis and promotes tumor progression in hepatocellular carcinoma. Oncotarget. 2015 May 10;6(13):10880-92

[4]https://www.proteinatlas.org/ENSG00000164961-WASHC5

ENSG00000070718.12	ENSG00000073921.18	ENSG00000076826.10	ENSG00000078177.14	ENSG00000085871.9	ENSG00000091136.15
ENSG00000095002.15	ENSG00000099203.7	ENSG00000099783.12	ENSG00000105976.16	ENSG00000111361.13	ENSG00000117500.13
ENSG00000132341.12	ENSG00000134250.20	ENSG00000136068.15	ENSG00000149136.9	ENSG00000151233.11	ENSG00000157404.16
ENSG00000158270.12	ENSG00000163249.13	ENSG00000164062.13	ENSG00000164331.10	ENSG00000164466.13	ENSG00000166598.15
ENSG00000169710.9	ENSG00000170365.10	ENSG00000170558.10	ENSG00000176171.11	ENSG00000196628.20	ENSG00000213551.7
ENSG00000229807.12	ENSG00000239264.9				

Table 4: List of female-male hubs in DEnetwork under cancer condition

- **TMED5** (ENSG00000117500.13) relevance in hepatocellular carcinoma was investigated in [5];

Instead, for the female population we have:

- **CAMSAP3** (ENSG00000076826.10), belonging to a wider family of genes studied in [6];
- **LAMB1** (ENSG00000091136.15), whose relevance for hepatocellular carcinoma progression was studied in [7];
- **MSH2** (ENSG00000095002.15), whose potentiality as a biomarker for hepatocellular carcinoma was investigated in [8];
- **ANKRA2** (ENSG00000164331.10), that has low cancer specificity but whose expression was detected in most of the patients suffering of liver cancer [9];
- **TCF4** (ENSG00000196628.20), whose relevance in hepatocellular carcinoma was investigated in [10].

However, none of this results is about specific gender-based investigation that might highlight sexual dimorphism in the hepatocellular liver carcinoma. This could either mean that there are no explicitely markers for differences between male and female population, or (way more likely) that our analysis needs refinement to target more relevant genes: for example, [11] effectively identifies some markers through Protein-Protein Interactions (PPI) models.

3.3 Patient Similarity Network (PSN)

We applied the Louvain algorithm to the PSN obtained on the cancer condition data merging back male and female tables. We retrieved the correlation matrix with a value for the threshold set to 0.85.

[5] Cheng X, Deng X, Zeng H, Zhou T, Li D, Zheng WV. Silencing of TMED5 inhibits proliferation, migration and invasion, and enhances apoptosis of hepatocellular carcinoma cells. Adv Clin Exp Med. 2023 Jun;32(6):677-688

[6] Wattanathamsan O, Pongrakhananon V. Emerging role of microtubule-associated proteins on cancer metastasis. Front Pharmacol. 2022 Sep 14;13:935493

[7] Liu T, Gan H, He S, Deng J, Hu X, Li L, Cai L, He J, Long H, Cai J, Li H, Zhang Q, Wang L, Chen F, Chen Y, Zhang H, Li J, Yang L, Liu Y, Yang JH, Kuang DM, Pang P, He H, Shan H. RNA Helicase DDX24 Stabilizes LAMB1 to Promote Hepatocellular Carcinoma Progression. Cancer Res. 2022 Sep 2;82(17):3074-3087.

[8] Hong S, Zhang J, Liu S, Jin Q, Li J, Xia A, Xu J. Protein profiles reveal MSH6/MSH2 as a potential biomarker for hepatocellular carcinoma with microvascular invasion. Hepatol Res. 2024 Feb;54(2):189-200

[9] https://www.proteinatlas.org/ENSG00000164331-ANKRA2/pathology

[10] Teng K, Wei S, Zhang C, Chen J, Chen J, Xiao K, Liu J, Dai M, Guan X, Yun J, Xie D. KIFC1 is activated by TCF-4 and promotes hepatocellular carcinoma pathogenesis by regulating HMGA1 transcriptional activity. J Exp Clin Cancer Res. 2019 Jul 24;38(1):329.

[11] Wu Y, Yao N, Feng Y, Tian Z, Yang Y, Zhao Y. Identification and characterization of sexual dimorphism-linked gene expression profile in hepatocellular carcinoma. Oncol Rep. 2019 Sep;42(3):937-952

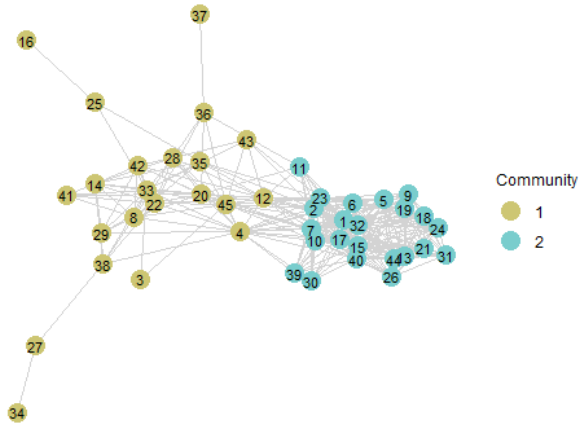

Figure 8: Patient similarity network highlighting the two communities retrieved by Louvain algorithm

The two communities may reflect some interesting patterns in the clinical records of the patients. Despite the fact that in both the communities the majority of the individuals are recorded as dead, we see that the first community is mostly populated by women, while the second one is mostly populated by men: this can be very interesting for further developments in sexual dimorphism in liver hepatocellular carcinoma. Moreover, we found out that in the second community most of the individuals have stage I cancer, while in the first one the highest percentage is represented by stage II or further cancer. We have summarised our findings in the following table:

	Alive	Dead	Stage I	Stage > II	Female	Male
Community 1	40.91 %	59.09 %	26.32 %	73.68 %	59.09 %	40.91 %
Community 2	21.74 %	73.91 %	55.56 %	44.44 %	26.09 %	73.91 %

Table 5: A table with some insights on the two communities (inconsinstencies might be due to misreporting)

4 Optional tasks

4.1 Differential co-expressed network comparing male and female population under normal tissue

We repeated the analysis we did in [3.2.2] comparing male and female population under normal tissue, so we kept the same threshold for the Z-scores and the same confidence level for the hubs.

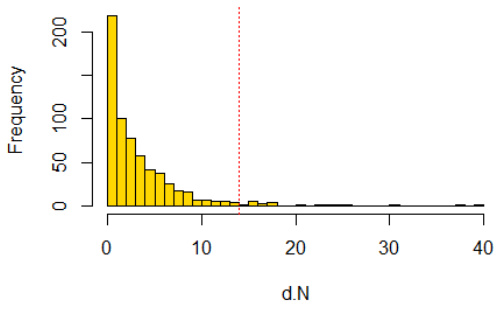

Figure 9: Normal tissue degree distribution;

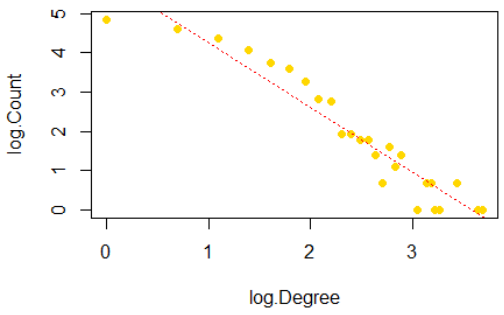

Figure 10: Normal tissue degree distribution in log-log scale;

We can already see that here a linear model better explains the log-log degree distribution: infact, the

maximum likelihood estimator for the gamma parameter in the power-law distribution functional is 2.98, which is reasonable for a scale-free network. Under this results, we expect to see fewer hubs: retrieving a similar number of hubs than before confirms that to assess the network structure properly we would need finer tools.

ENSG00000007350.17	ENSG00000048649.13	ENSG00000067048.17	ENSG00000083312.17	ENSG00000090612.22	ENSG00000091879.14
ENSG00000101972.19	ENSG00000105821.15	ENSG00000115392.13	ENSG00000116984.14	ENSG00000117155.17	ENSG00000120605.16
ENSG00000125965.9	ENSG00000136379.12	ENSG00000136935.14	ENSG00000138078.16	ENSG00000138160.7	ENSG00000148943.12
ENSG00000158270.12	ENSG00000162711.18	ENSG00000164305.19	ENSG00000164828.18	ENSG00000164867.11	ENSG00000165169.11
ENSG00000170558.10	ENSG00000185591.10	ENSG00000198087.7	ENSG00000229807.12		

Table 6: List of DEGs overlapping between men and women data under normal tissue condition

(a) Male DE network

(b) Female DE network

(c) Cancer tissue DE network

(d) Normal tissue DE network

Figure 11: Visualizing differential expression network and the subnetworks induced by the retrieved hubs

4.2 Signed network analysis

We now want to build a signed network from the one we built in [3.2.2] such that

$$A_{i,j} = \mathbb{1}(Z_{i,j} \geq \overline{Z}) - \mathbb{1}(Z_{i,j} \leq \overline{Z}) \tag{3}$$

From such network we can extract two subnetworks: one with only positive edges, the other with only negative edges. We claim that signed networks typically are used to represent the nature of the interactions, such as synergistic/antagonistic in biological systems (or also friendly/hostile in social networks), so consider hubs with highest positive and negative degree values means examining not only the graph connectivity but also the type of interactions among nodes.

The following figure illustrate the two subnetworks:

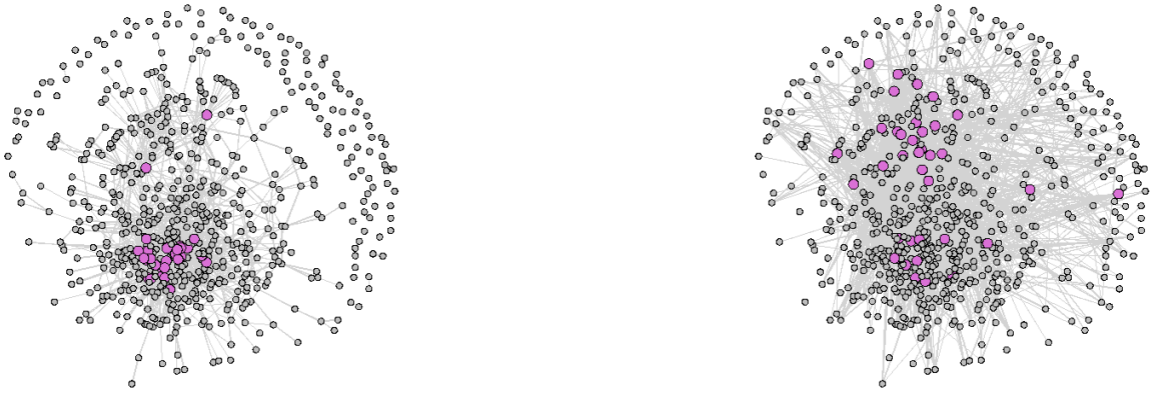

Figure 12: Positive (left) and negative (right) signed network;

ENSG00000070718.12	ENSG00000136068.15	ENSG00000151233.11	ENSG00000169710.9
ENSG00000170558.10	ENSG00000213551.7	ENSG00000229807.12	

Table 7: Hubs overlapping between positive and negative subnetworks

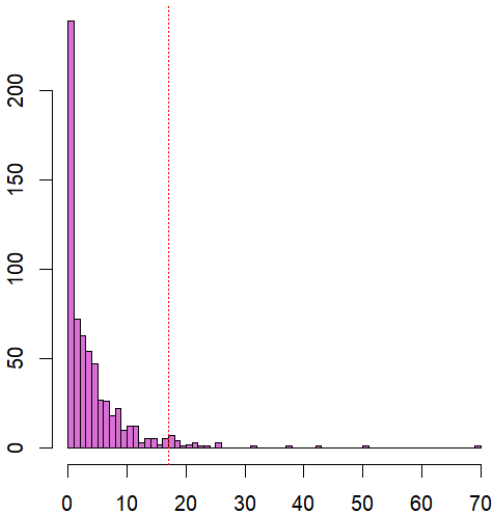

Figure 13: Degree distribution in the positive subnetwork;

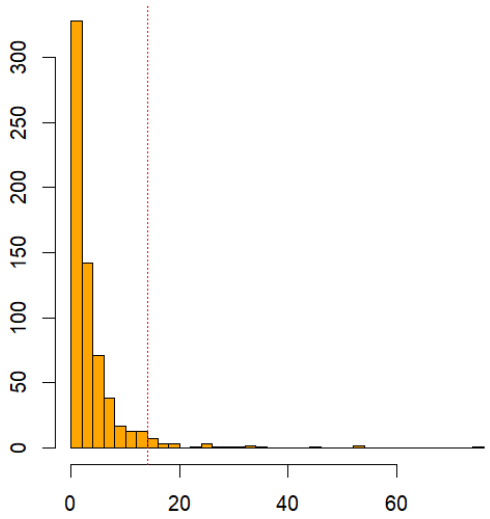

Figure 14: Degree distribution in the negative subnetwork;

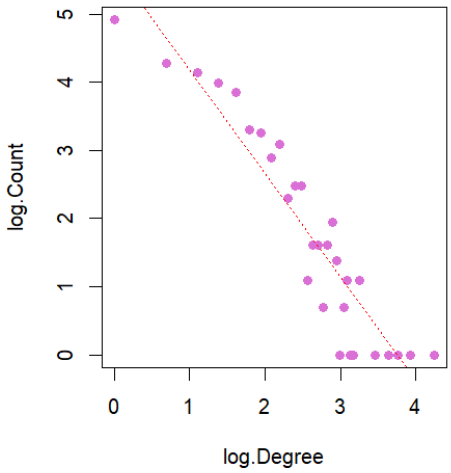

Figure 15: Degree distribution in the positive subnetwork;

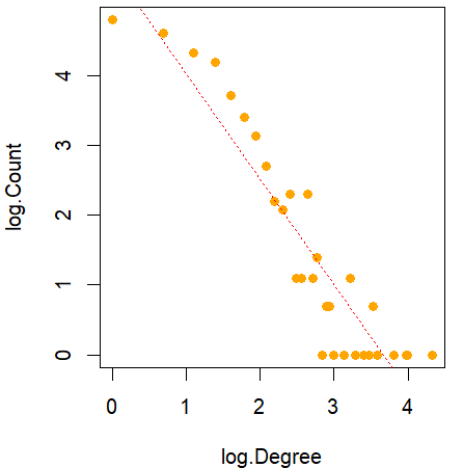

Figure 16: Degree distribution in the negative subnetwork;

Looking again at the network structure, the two both seem to show scale-free behaviour when looking at the plotted degree distribution: the MLE for the γ parameter in the functional of the distribution is 3.10 for the positive subnetwork and 2.56 for the negative subnetwork. We conclude that the latter definitely have scale-free structure, while the former should better assessed. This also explains why the positive subnetwork is very dense in between hubs, while the negative one spreads also towards more peripheral nodes.