Differential Analyses of Kidney Gene Expression

D'Erasmo Giulio
Sapienza, university
of Rome, Italy

Potì Andrea
Sapienza, university
of Rome, Italy

Mehrdad Hassanzadeh
Sapienza, university
of Rome, Italy

Abstract

The topic of the analysis is focused on making a study about a subcategory of RCC (Renal Cell Carcinoma) called KICH, also known in the literature as chromophobe renal cell carcinoma (ChRCC). The analysis was conducted extracting qualitative and quantitative insights from The Cancer Genome Atlas Kidney Chromophobe Collection (TCGA-KICH) data, using network science techniques, in order to identify the genes potentially responsible of cancer. The findings could potentially help doctors detect the cancer earlier. We discover that the differential genes in Cancer tissue which are significant correlated have high positive correlation value, without having switching of genes. Also we found the recurrence of the genes "PTH1R" during the analysis which is know to be expressed in chrcc tissue. Despite the result we didn't found any novel literature about our data and findings, due to the rare condition in exam.

1 Introduction

Human RCC (Renal Cell Carcinoma) is categorized in subcategories based on its appearance under a microscope, which includes chromophobe (KICH), clear cell (KIRC), papillary (KIRP), collecting duct, and unclassified RCC. The most prevalent subtype is KIRC which represents 75–80% of RCC cases, whereas KIRP (10–15%), KICH (5%) and rare RCCs comprise the remainder. In our analysis we focus on KICH, also known in literature as Chromophobe renal cell carcinoma (ChRCC), which is originates from intercalated cells in the distal convoluted tubules of the nephron and is characterized by a relatively good prognosis and exhibits a low degree of malignancy (1). In order to study the genes that characterize the disease we identify the differentially expressed genes (DEGs) through statistical analysis of the transcriptome data. Those are used to build co-expression networks and differential co-expressed network which are able to give insight regarding hubs (genes with high number of degree). In the end we build a Patient Similarity Network (PSN) using clinical data in order to characterize similar patients. As bonus of our study we perform the study varying similarity measure or the type of threshold discussing the difference or similarity of the results.

2 Materials and Methods

2.1 Data collections

We analyzed gene expression data from GDC Data Portal selecting the **Kidney Chromophobe Project with ID: TCGA-KICH** only for whom cancer and normal tissue files are available. We use the library *TCGAbiolinks* to download the data directly with R. In order to deal with this dataset, we extract only patients for which we have one sample and are common to both the type of files having a total of 24 patients. The smaller number of patient is due to the rare type of RCC we are studying. As final pre-processing step we ensure to have no missing data and to remove genes which have zero values. We will work with 18423 genes. More details are shown in the Table 1 below.

	Cancer files	**Normal Tissue files**
n.patients	65	25
no zero genes	18886	22863
missing data	No	No

Table 1: TCGA-KICH data information.

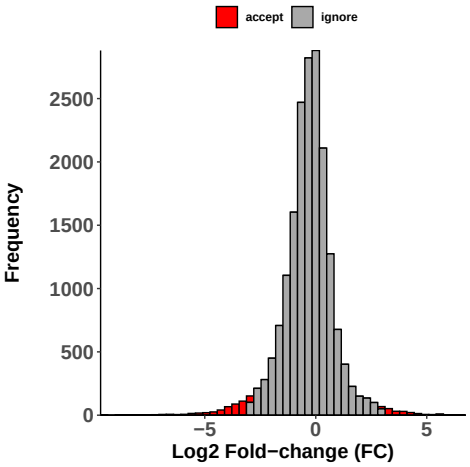

Figure 1: FC histogram. Grey bars represent the RNAs discarded, red bars the retained ones.

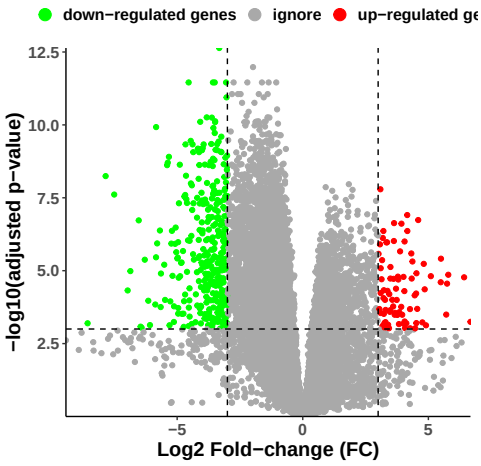

Figure 2: Volcano Plot. Grey points represent discarded according to the selected threshold, green points are the down-regulated RNAs, red points are the up-regulated RNAs.

2.2 Differentially Expressed Genes (DEGs)

In order to analyze our data we need to select the genes that are differentialy expressed, meaning that a difference or change observed in read counts or expression levels/index between two experimental conditions is statistically significant. Hence we start with a filtering phase that aims to select the genes that are varying on average a lot and in a statistically significant way between the two conditions, namely we will look to Fold-Change (FC) measure and Student's t-test.

Fold-Change: The Fold change indicates whether a gene is up-regulated (more expressed) or down-regulated (less expressed) in a group type with respect to another group type. In our case Cancer vs Normal types of tissue. In detail we use the **log2FC** which model proportional changes rather than additive changes, which are biologically more relevant.

$$\text{FC} = \log_2 \frac{\langle \text{Cancer data} \rangle}{\langle \text{Normal data} \rangle}$$

Student's t-test: Due to the limitation of the FC: many false positives; resulting low-expressed genes differentially expressed with non-significant variations; we need to adopt a paired Student's t-test, which perform a comparison using the mean of the data, in order to filter by the corresponding p-value. Here the null hypothesis H_0 is that the difference is due to chance, while the alternative hypothesis H_1 is that the difference indeed exist. If the p-value of the test that measure the statistically significant of the comparison is lower than a certain threshold, we will reject the null. We also account the problem of multiplicity hypothesis testing adjusting the p-value according to the **False Discovery Rate** method, which is the least stringent of all corrections and provides a good balance between discovery of statistically significant genes and limitation of false positive occurrences.

In order to conclude the filtering phase we need to select a threshold that will allow to have only a subset of hundreds of genes. We end up with $|\rho| \leq 0.001$ and $|\text{FC}| > 3$. In Figures 1, 2 we represent the result of our choices having a final number of genes of 427, keeping 2% of the total genes. In the first figure there is the histogram of Log2FC, showing how many samples fall in the corresponding bins and highlighting in red the bins we accept after the tresholding cut. After that we represent the Volcano Plot which underline the genes that are statistically significant with green dots, down-regulated genes,and with red dots, the up-regulated genes.

2.3 Co-expression networks

Now we are ready to build a co-expression networks and start our firsts analysis. A 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 **signif-**

icant co-expression relationship between them. We will explore the co-expression relationship using **Pearson** correlation coefficient connecting with a link two genes if the measure is above a certain threshold. In order to keep only the significant link we perform a correlation test on them (adjusting for multiplicity with FDR correction). We set no link if $|\rho| < 0.65$ having a trade-off between a small number of links in order to have a manageable network (high threshold) and the number of connected components should be as small as possible in order to preserve the integrity of the network (small threshold). In practice we use the function *cor* and *corr.p* from the packages *Stats* and *psych* respectively for building the adjacency matrix and than we use *Igraph* in order to build our networks and analyze them.

The first thing we check is if the degree distribution follows a **power law**: $p_k \sim k^{\gamma}$, where k is the degree of the nodes, in order to ensure that we are dealing with a **scale-free network**. We use the package *poweRlaw* in R which perform a **Kolmogorov-Smirnov** statistical test in order to see if our data fit a power law distribution, giving us the corresponding statistical value and p-value.

If the networks are indeed scale-free, we look at their **hubs**, which are the 5% of the nodes with highest degree values, and compare them.

2.4 Differential Co-expressed Network

Instead of establishing that the co-expression is significant in one condition and not in the other, one could test directly if the change in co-expression is significant. Differential networks encode the changes in connections among nodes between the conditions or states. In order to build the network we compute the **Z-score**:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

where z_1 and z_2 are the Fisher z-transformation of the DEGs, obtained using the function *fisherz* from the *psych* package in R, and n_1, n_2 are the number of patients. Setting no link if $|Z| < 4$, we build the adjacency matrix of the differential co-expression graph.
We perform than the same analysis of the co-expression networks, adding a comparison between the hubs genes we found.

2.5 Patient Similarity Network (PSN)

At last point we want to build a Patient Similarity Network (PSN) where each node is an individual patient and an edge between two patients corresponds to **pairwise similarity** for a given feature. In this paradigm, each input patient data feature (e.g., age, sex, mutation status) is represented as a network of pairwise patient similarities.
Each feature is represented as a different "view" of patient similarity that can be integrated with all the other views to identify patient subgroups or predict outcome. In our case we study two different type of PSN: one using the clinical data, an other using cancer genomics data (gene expression data), and we look if arise community or different pattern.

3 Results and Discussions

3.1 Co-expression networks

Starting from our DEGs, we study the correlation matrix of our genes to build the adjacency matrix of the co-expression networks. What we find interesting is that in the Cancer Networks the accepted genes, the ones that are highly correlated, are **all positive correlated**, meaning that expression of one gene increases with the increase in the expression of its co-expressed gene. An high proportion of genes have a low correlation around zero and are discarded. Similar results are found in the Normal Networks but having here a small number of highly negative correlated genes that pass the threshold (Figure 3, 4).

Having our networks we can display the **degree distribution** of those. Figure 5 and 6 shows the plot in exam in log-log axis of the cancer and normal network namely. It's already evident the scale-free property if we focus on the hubs which occur with lower probabilities for high degree nodes. Performing instead a more rigorous numerical analysis,as said before, we obtain indeed that **both the graphs follow a power law distribution**. Results are shown in Table 2. Both the distribution get an high p-value, so we accept

	γ	**KS**	**p.value**
cancer net	2.49	0.08	0.99
normal net	2.46	0.07	0.84

Table 2: Scale free fitting of the co-expression networks.

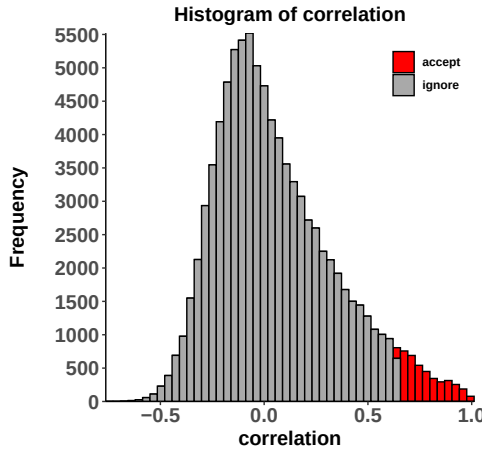

Figure 3: Histogram of the correlation for the cancer DEGs. The red bar are the one that satisfy the threshold for building of the co-expression network

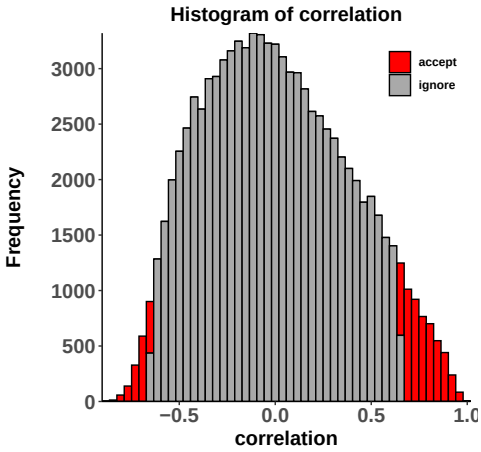

Figure 4: Histogram of the correlation for the normal tissue DEGs. The red bar are the one that satisfy the threshold for building of the co-expression network.

the null hypothesis that the original data could have been drawn from the fitted power-law distribution. Also the estimated exponent of the power law is in $(2, 3)$ which is a typical range for scale-free net.

Table 3 shows the names of the **hubs** of the networks. Normal tissue net has more hubs than Cancer network having only "PTH1R" "MYL3" has common genes.

Cancer net		Normal net	
ABCB1	KL	MGST1	FMO1
GSTM3	KCP	BST1	NRXN2
KCNJ1	KCNJ16	HSPA2	AIF1L
PTH1R	MYL3	EHF	GIPC2
ADAMTS9-AS1	TMEM169	GIPC2	ENPEP
IQUB	SLC16A4	RBP5	AMDHD1
MYO7B	PDZK1	SUSD4	SERP2
KCNJ10	AFMID	LGI2	BMP6
LRRC19	CYS1	ENPP3	ALDH4A1
C21orf62	LINC02532	PTH1R	MYL3
		AK4	CFI
		EXOC3L4	TENM2-AS1

Table 3: List of the hubs for the different networks.

Let's investigate now only the cancer related hubs. Looking at the subgraphs (Figure 7) we found that the genes with high beetweeness centrality is "AFMID", which indeed connects the two subnetworks in the graph; while the ones with high centrality are "ABCB1", "MGST1", "EHD3". Due to the relative small sample of genes we take for the analysis (due to having to use strong threshold for Fold Change and p-value) and the fact that there are few analysis in the online literature related to the KICH dataset, we didn't find anything worth to cite related to our hubs gene, except they are related (known to be) in the cancer tissue of the kidney.

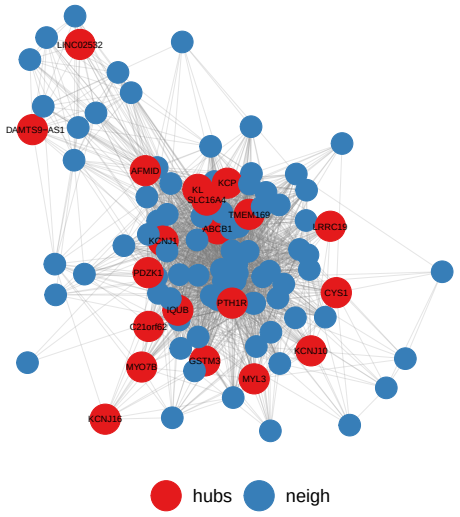

Figure 7: Only cancer hubs subnetworks. In red are highlighted the cancer genes hubs.

3.2 Differential Co-expressed Network

Looking at the degree distribution, we test if it follows a power low distribution and it does, as shown in Figure 8, and conformed by a statistical test having p-value $= 0.99$ and KS-stat $= 0.06$,

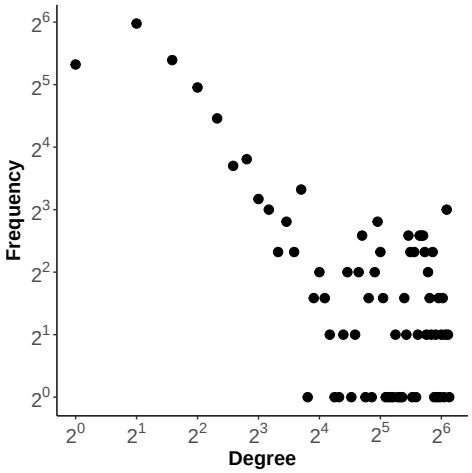

Figure 5: Log-Log Degree distribution of the cancer network.

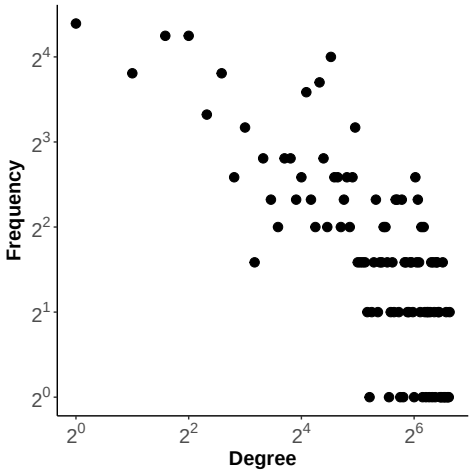

Figure 6: Log-Log Degree distribution of the normal network.

hence confirming the null H_0. We can now that study the hubs.

There are 22 hubs in total, three of those with the higher degree are: ”MGST1”, ”FMO1”, ”ABCB1”, ”NCS1” and ”NRXN2”. The ones in common with the co-expression networks of Cancer and Normal are only two: ”PTH1R”, ”MYL3”.

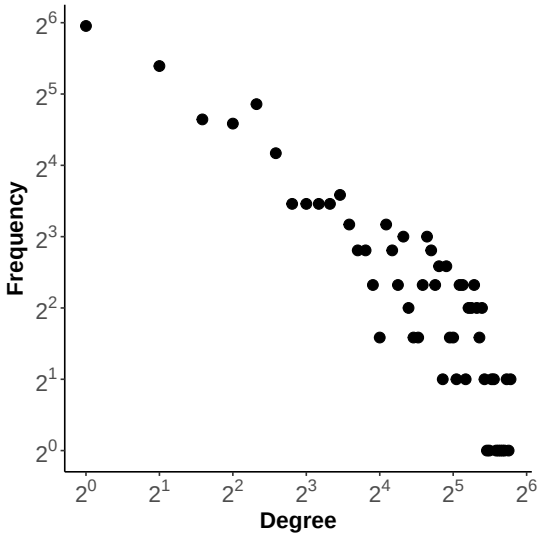

Figure 8: Degree distribution of differential co-expression network.

3.3 Patient Similarity Network (PSN)

First of all we compute the PSN with the cancer gene expression data. We use the function *cor* in order to build the Pearson correlation adjacency matrix. What we found is that at genome level all the patients were highly positive correlated, and due to the small number, they reduced to be in the same cluster which was performed using the *Louvain algorithm* of the *Igraph* package in R. We decide so to do another analysis looking at the clinical data. After a brief pre-process of keeping the interesting feature and be sure to remove the ones with all missing data, we build our PSN using the following variables: ”ajcc pathologic stage”, ”prior malignancy”, ”ajcc staging system edition”, ”ajcc pathologic t”, ”ajcc pathologic n”, ”race”, ”gender”, ”treatments pharmaceutical treatment or therapy”. We obtain using the Louvain algorithm 3 clusters which are shown in Figure 9. Digging inside the clusters we notice that what really matters was the difference in ”ajcc pathologic n” and ”ajcc pathologic t”.

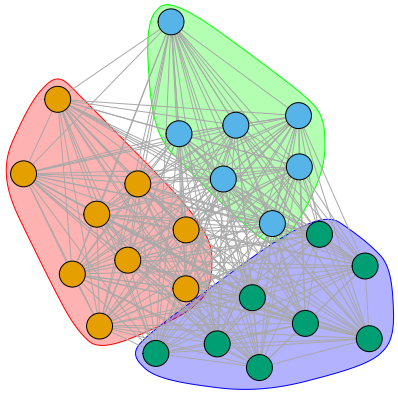

Figure 9: PSN of clinical data.

	Hard Tresholding Cancer Network	Hard Tresholding Normal Network	Soft Tresholding Cancer Network	Soft Tresholding Normal Network
Global efficiency	0.412	0.452	0.211	0.380
Avg local efficiency	0.735	0.793	0.525	0.677
Diameter	8	6	17	9
Max degree (Normalized index)	0.291	0.281	0.183	0.293
Number of connected components	1	2	51	22
Avg Closeness Centrality	0.661	0.743	0.040	0.612
Avg Betweeness Centrality	0.017	0.040	0.059	0.023
Avg Hub score	0.245	0.219	0.153	0.210

Table 4: Summary of topology index related to the networks obtained using Hard-tresholding and Soft-tresholding both to the Cancer and Normal tissue network.

4 Bonus

4.1 Hard-thresholding and Soft-thresholding

Here we are interested in comparing the network topology obtained applying hard-thresholding and soft-thresholding. The difference between the two type of tresholding is:

- in **Hard-thresholding** the correlation coefficient $s_{i,j}$ between gene i and j is assessed with respect to a threshold τ and both genes are connected by an edge:

$$A_{ij} = \begin{cases} 1 & \text{if } s_{i,j} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

  The values of τ was set to 0.5;

- in **Soft-thresholding** each entry of the adjacency matrix is evaluated using an adjacency function, the most used ones are *sigmoid* and power functions. In our analysis we decided to choose the sigmoid:

$$\begin{aligned} A_{ij} &= \text{sigmoid}(s_{i,j}, \alpha, \tau_0) \\ &= \frac{1}{1 + e^{-\alpha(s_{i,j} - \tau_0)}} \end{aligned} \tag{2}$$

  The values of α and τ_0 was set both to 0.1.

After created the networks out from these two new adjacency matrices, we computed a series of useful indexes and values to compare the different topologies. The new networks are 4 in total, because we have cancer and normal condition and for each of the two we applied hard and soft thresholding separately.

We sum up all the information in Table 4 in which each row is a different network and the columns contain the comparison values. The statistics are made in R using the functions available in *Igraph*.

4.2 Overlap of hubs with a different centrality index

Now we recompute using a different centrality index, the **betweeness centrality**, the 5% of the nodes with highest CI values and compare the result with the degree-based hubs.
We obtain that the hub genes in common are "AFMID" with respect the Cancer Network and "AK4" with respect the Normal Tissue network. As we analyze in the section 3.1 we retrieve the "AFMID" genes which was a degree hubs with highest betweeness centrality. While the latter gene doesn't come out as particular genes before despite being a degree hubs.

4.3 Co-expression study using different similarity measure

We perform a part of the whole study did in the section 2 and 3, but this time using as initial correlation method two different similarity scores. We used **Spearman** and **Kendall** correlation methods to create the co-expression networks and we find the common hubs between the 2 conditions with these new measures.
The result is that the genes stay the same with all the different methods and they are "PTH1R" and "MYL3".

We than try using a totally different type of correlation measure which is the **biweight midcorrelation**, which is considered to be a good alterna-

tive to Pearson correlation since it is more robust to outliers. Also we found in literature that this metrics outperform the Mutual Information ones (2). We start than building the adjacency matrix using the *bicorAndPvalue* functions provided by the package *WGCNA* in R, adjusting also for multiplicity using FDR method. We notice at first that now appears negative links inside the Cancer adjacency matrix with respect the absence in the standard analysis. Than we fit the degree distribution of the two networks with respect the power law function, having good results for the Normal Tissue networks and the Cancer one. Despite the change in the networks link we found that also here the common hub gene "PTH1R".

4.4 PSN using normal condition and comparison of the community structure with cancer one

Here we applied the same analysis performed in section 3.3 but using gene expression profiles related to normal condition and the result obtained is shown in Figure 10. As what we have obtained in the community detection for the Cancer genomic data, we found one single community due to the high correlation matrix and the smaller sample of patients.

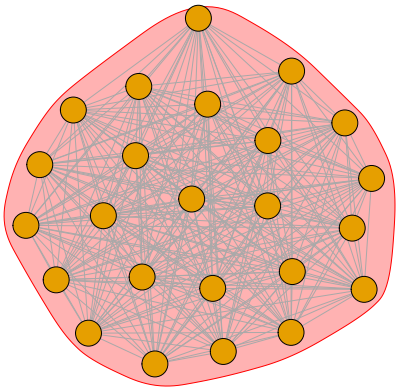

Figure 10: PSN of Normal tissue gene expression data.

References

[1] Badowska-Kozakiewicz AM, Budzik MP, Koczkodaj P, Przybylski J. *Selected tumor markers in the routine diagnosis of chromophobe renal cell carcinoma.* Archives of Medical Science. 2016;12(4):856-863. doi:10.5114/aoms.2015.51188.

[2] Song, L., Langfelder, P. Horvath, S. *Comparison of co-expression measures: mutual information, correlation, and model based indices.* BMC Bioinformatics 13, 328 (2012). https://doi.org/10.1186/1471-2105-13-328