DIGITAL EPIDEMIOLOGY AND PRECISION MEDICINE: ANALYSIS OF PRAD HUB GENES

LEONARDO LAVAGNA

Abstract. Prostate Adenocarcinoma (PRAD) is a very common malignant tumor that affects men primarily in their old age, which incidence has been appreciably increasing over the last few decades. The aim of this study is to identify the hub genes closely related to the PRAD. The study started by considering counts informations about 18000 genes analyzed on 500 patients (taken from the Genomic Data Commons Data Portal, [TCGA-PRAD]), from which the Differentially Expressed Genes (DEGs) were extracted and their regulation studied. Once we obtained the DEGs, we performed analysis on both Coexpression Networks and Differential Coexpression Network, and several hubs of genes have been extracted. We then worked on Patient Similarity Networks and we extracted communities of similar patients affected by the tumor. In conclusion we compared our results with the available medical literature. Many of our statistical findings are in accordance with the literature, but some new genetic information resulted statistically significant and may be of interest for further studies.

Contents

Introduction

Prostate Adenocarcinoma (PRAD) is one of the most common cancers affecting men in their late adulthood (cfr. [Rawla et al 2019]). It originates in the prostate's tissue and it is often treatable (cfr. [Kaler et al 2020]). PRAD is mainly divided into 3 categories, including Adenomas and Adenocarcinomas, Cystic, Mucinous and Serous Neoplasms, Ductal and Lobular Neoplasms as reported in [TCGA-PRAD]. As a result of early diagnosis, the mortality rate of PRAD fall, although the incidence of PRAD continues to rise (cfr. [Potosky et al 1995], [Brawley 1997] and [MaryBeth et al 2020]). The high incidence of prostate cancer depends on many uncontrollable factors, such as specific genetic mutations, traumatic events, diet and lifestyle. It is therefore very important to identify those hub genes involved in PRAD to help develop new therapies and early screening tests. In the following work, a multistep analysis has been conducted:

- extrapolation and preprocessing the relevant data from [TCGA-PRAD];
- construction of Differentially Expressed Genes (DEGs);
- construction of Coexpression Networks (exploiting hard-thresholding and comparing it with soft-thresholding), and identification of candidate hub genes by considering different centrality measures;
- analysis (with different similarity measures) of Differential Coexpressed Network based on DEGs;
- extrapolation of communities (with different methods and via clustering techniques) in Patient Similarity Networks.

When possible our findings were compared with available medical literature.

1. Materials and Methods

In this section we explain all the processing steps and analysis performed. All the code we've written to carry out the project can be accessed from GitHub, see [GitHub TCGA-PRAD].

1.1. **Set Up, Data extrapolation and processing.** Throughout the project we used the `R` programming suite for statistical computing in conjunction with the Cancer Genome Atlas, that is the main platform where we can retrieve a wide collection of genomic data about 60660 genes and 500 patients. For our analysis we handled two datasets extracted from TCGA: the tumor tissue and the normal tissue. The libraries `TCGAbiolinks` and `SummarizedExperiment` were used to run queries on the Atlas web platform and get the relevant data[1]. The datasets are given in tabular form and are composed by patients in the columns and the genes in the rows. The values on each cell are raw counts of transcriptome profiling of each gene (`HTSeqCounts`). The gene expression of each gene needed to be assessed since it could be used to remove the least significant data from the massive datasets we retrieved. Moreover there could be redundant data, that is why we used the `dplyr` and `stringr` libraries to selected common patients (through binary masks and a list of participants to match their original barcode identifier) and common genes through the two kinds of datasets. After getting all the data from TCGA, after down-sizing the retrieved datasets by removing least significant records and after

[1] In the notebook in [GitHub TCGA-PRAD] the relevant data is organised and split into different tables: `rna_expr_data_C`, `rna_gene_data_C`, etc...

removing redundant entries we obtained two datasets, called `rna_expr_data_C` (Tumor) and `rna_expr_data_N` (Normal) that we used throughout the project.

1.2. **Differentially Expressed Genes (DEGs).** To perform this task we used the `DESeq2` package, that expects count data in the form of a matrix of integer values where each entry-value v_{ij} corresponds to the number of read assignable to gene i in sample j. This method is based on a negative binomial distribution, can perform multiple comparisons and output a flexible object called `DESeqDataSet`. By running the `DESeqDataSetFromMatrix` function[2] on the relevant data given in Subsection 1.1 we build a `DESeqDataSet`. From it we removed lowly expressed genes (keeping only genes that have at least 10 reads total, where 10 is a first crass trashold). The standard differential expression analysis (DEG Analysis) is given by running a single function[3] `DESeq`. From the DEG Analysis' output, we retrieved only those results having an adjusted $p-$value below a given FDR cutoff $\alpha = 0,05$, and an $|LFC| \geq 1,2$ (cfr. Figure 1).

1.3. **Coexpression networks.** Once the differentially gene expression analysis was concluded, our goal was to define two Gene Coexpression Networks (GCNs), with respect to the normal condition and the tumor condition. Coexpression networks are performed using two expression matrices[4] after being filtered as explained earlier. From the two gene expression matrices, Pearson's correlations between every pair of genes were computed[5]. The two new correlation matrices are then transformed into binary correlation matrices[6] from which we got the two Gene Coexpression Networks, the associated Degree Distributions and the Hubs (5% of nodes with highest degree values). The main results are plotted in Figure 2, Figure 4 and Figure 5. For completeness alternative correlation measures were also studied, all the details can be found in the notebook in [GitHub TCGA-PRAD] and the main results are summarized in Appendix B.

1.4. **Differential Coexpression Network.** The Differential Coexpression Network (DCN) was performed using the same two correlation matrices obtained during the contruction of GCNs. Fisher transformation was applied to convert the two correlation matrices into Z-Score matrixes. From the two Z-score matrices associated to the two conditions (Tumor and Normal) an overall Z- Score was obtained to assess the differential correlation. This correlation is given by the formula

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

[2]In the notebook in [GitHub TCGA-PRAD] the main input of this function is given by `PartecipantsCondition` which is obtained by merging `rna_expr_data_C` with `rna_expr_data_N`.

[3]The function takes as input our filtered `DESeqDataset`

[4]One matrix for Tumoral samples and the other one for Normal samples. In the notebook in [GitHub TCGA-PRAD] they are `rna_expr_data_C` and `rna_expr_data_N` as obtained after some preprocessing steps from the original datasets in Subsection 1.1.

[5]This step was preceded by a log-transformation of the data for technical reasons. In the notebook in [GitHub TCGA-PRAD] those correlation matrixes are `co_net_corr_dataC` and `co_net_corr_dataN`.

[6]See Appendix A where it is explained how to set a good (hard) threshold that allow to transform the correlation matrices into binary correlation matrices. In the notebook in [GitHub TCGA-PRAD] those are `co_net_corrBinary_dataC` and `co_net_corrBinary_dataN`

where n_1 and n_2 represent the sample size for each of the two conditions (Normal and Tumor). Once again a good correlation threshold must be found in order to have a network that is not composed of too many disconnected components, nor it is too dense. This technical point is explained in Appendix A, A.2. Once a good threshold has been found the same operations carried out to build the GCNs were also applied here to get a DCN.

1.5. **Patient Similarity Networks.** In a Patient Similarity Network (PSN) patients are clustered or classified based on their similarities in various features, including genomic profiles. We studied two kinds of PSNs: a first PSN obtained using clustering based on the tumor data extracted in Subsection 1.1 and Jaccard Similarity, a second PSN obtained using the Louvain method for community extraction from the correlation matrix obtained in Subsection 1.3 (see also Figure 6 and Figure 7).

2. Results and Discussion

In this section we elaborate on the results we obtained.

2.1. **Fold Change and Volcano Plot.** From the constructed DEGs we got the Volcano plot, a useful type of scatter plot that shows statistical significance (adjusted $p-$value) vs magnitude of change (Fold Change).

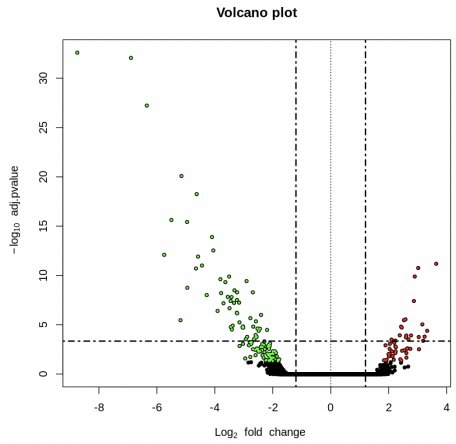


Figure 1. Volcano plot. Adjusted p-value $\leq 0,05$, $|LFC| \geq 1,2$

The genes highlighted in red are the most statistically significant up-regulated genes, the green ones are instead the most statistically significant down-regulated genes.

2.2. **Trasholds, Correlation in the Networks and Degree Distribution.** By setting an hard-threshold $\tau = 0.4$ (see Appendix A, A.2 for all the technical details) we have better performance in terms of network connectivity (i.e. the resulting networks remain intact) the correlation matrixes in the coexpression network are such that the correlation entries satisfy $|\varrho_{ij}| \geq 0.4$. Likewise with a Z-threshold $\tau_Z = 5$ (see Appendix A, A.2 for all the technical details) we have once again better

performance in the case of DCNs. With those thresholds we have the following plots in the case of GCNs.

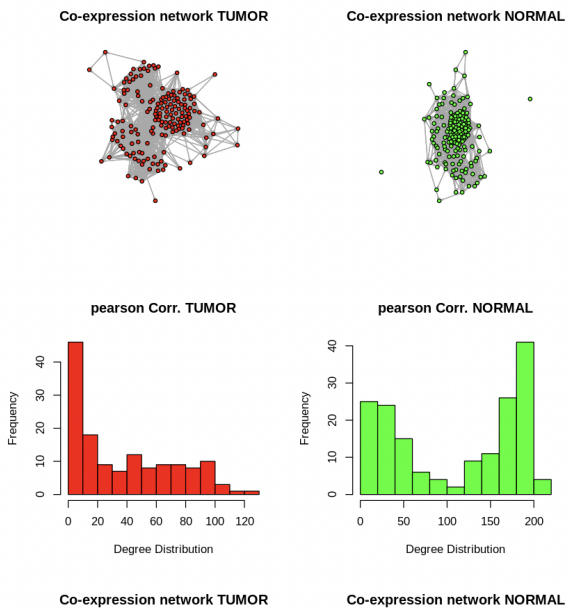


FIGURE 2. Coexpression Networks with $\tau = 0.4$ and respective Degree Distributions

The following plots instead are those about the DCNs.

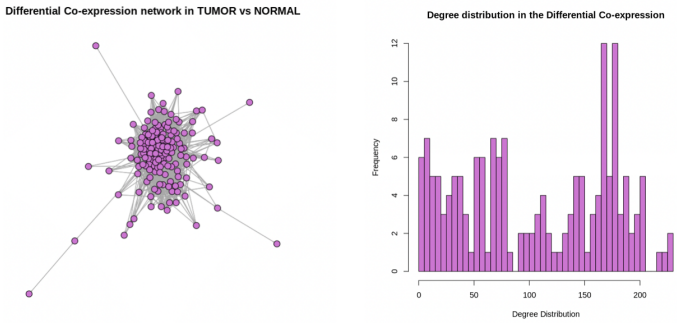


FIGURE 3. Differential Coexpression Network with $\tau_Z = 5$ and respective Degree Distributions

For completeness alternative correlation measures and different centrality measures were also studied, all the relevant plots can be found in the notebook in [GitHub TCGA-PRAD] and the main results are summarized in Appendix B.

2.3. **Candidate Hubs of Genes.** The hubs in the tumor and normal coexpression network can be deduced from the degree distribution plots as shown below in purple.

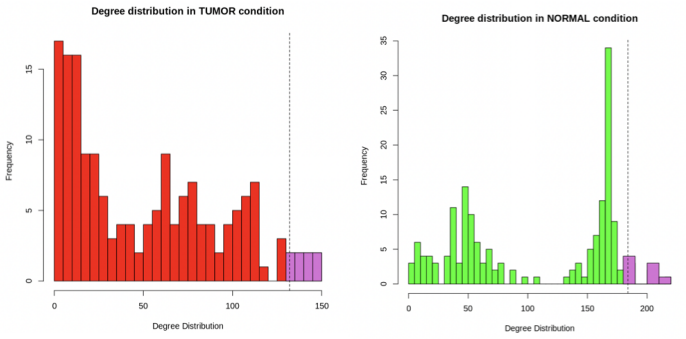


FIGURE 4. Hubs of genes from overall degree distribution

The hubs are

- tumor case: `ENSG00000176928.7`, `ENSG00000099957.16`, `ENSG00000128266.9`, `ENSG00000103485.19`, `ENSG00000120885.22`, `ENSG00000152137.8`, `ENSG00000109846.9`, `ENSG00000244509.4`
- normal case: `ENSG00000100505.13`, `ENSG00000147234.10`, `ENSG00000128266.9`, `ENSG00000048540.15`, `ENSG00000118298.12`, `ENSG00000087258.16`, `ENSG00000125257.16`, `ENSG00000136840.19`.

The only common gene is `ENSG00000128266.9`. Their degree distribution is as follows.

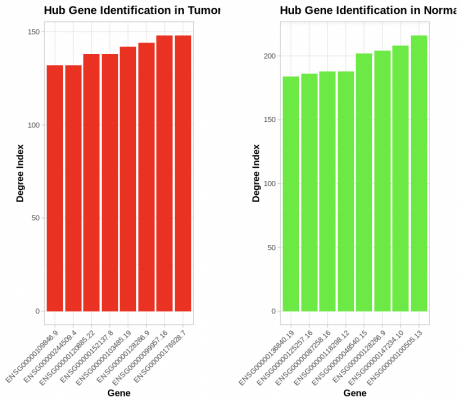


FIGURE 5. Hubs of genes (Normal and Tumor)

It is interesting to note that some of the hub genes found were also found in literature as explained in [Cristofani et al 2021] and [Brawley 1997]. See [AGO] for further references about the other hub genes we found.

2.4. **Communities and Clustering in Patient Similarity Networks.** We obtained clusters of patients using Jaccard similarity (Figure 6 left) and one community using the Louvian method (Figure 6 right). The most interesting case is the community detection since it is based on genomic data[7].

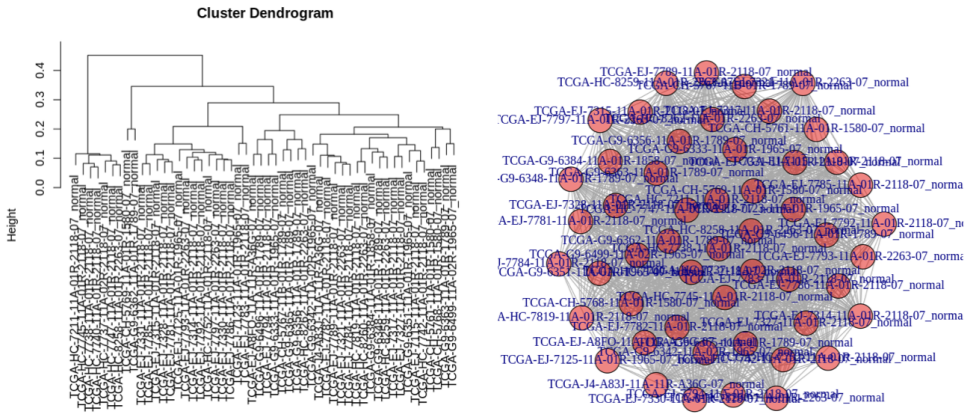


FIGURE 6. Jaccard clustering and community detection

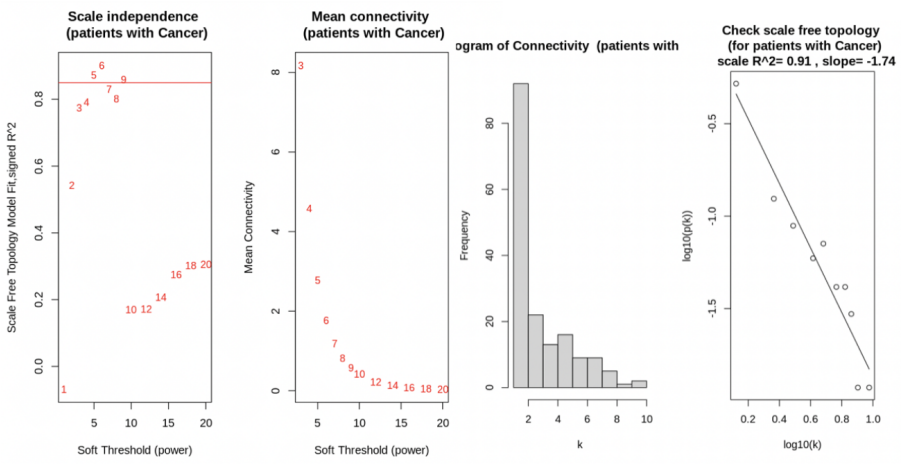
3. Conclusions

We identified two hubs of genes that have also been found in the specialized literature. Those genes are `ENSG00000120885.22` and `ENSG00000152137.8`. We've studied Gene Coexpression Networks and Differential Coexpression Networks using different similarity measures and different correlation measures. In each on these analysis the hub genes previously mentioned were always present.

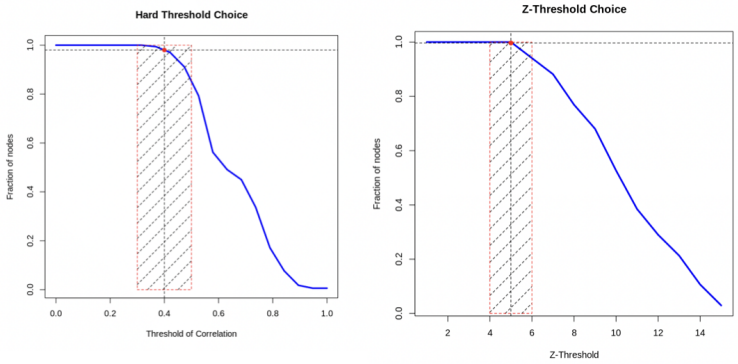
[7] In the notebook in [GitHub TCGA-PRAD] we used as genomic data the `co_net_corrBinary_dataC` studied in Subsection 1.3.

Appendix A

A.1 Soft Thresholding. Soft-thresholding assigns a connection weight to each gene pair in such a way that all the nodes of the network are ranked according to their connection strength. Here we considered power adjacency functions $a_{ij} = \varrho_{ij}^{\beta}$. By using the `WGCNA` package in `R`, the power $\beta = 6$ was selected in the tumor's case as the soft-thresholding to ensure a scale-free network as shown in the following figures.



A.2 Hard Thresholding and Hard-ZThresholding. Hard thresholding is the process of setting to zero the correlation coefficients whose absolute values are lower than the threshold τ, setting the to one otherwise. In order to find an optimal way to attribute the good hard threshold for our networks we created a set of functions (in the notebook in [GitHub TCGA-PRAD] those are `fractionNodes` and `OptimalThresholding`) that, within a range of possible thresholds, makes possible to find the optimal one that allows to have the main networks intact. This can be done, for example, by inspection of the following plots.

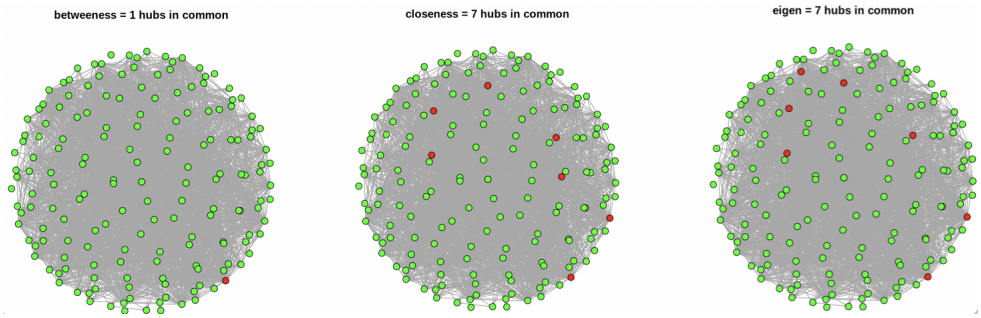


Appendix B

B.1 Results obtained using different centrality measures. After finding the hub genes in each corresponding coexpression network, we computed alternative centrality indices. We obtained the following hub genes in the tumor case:

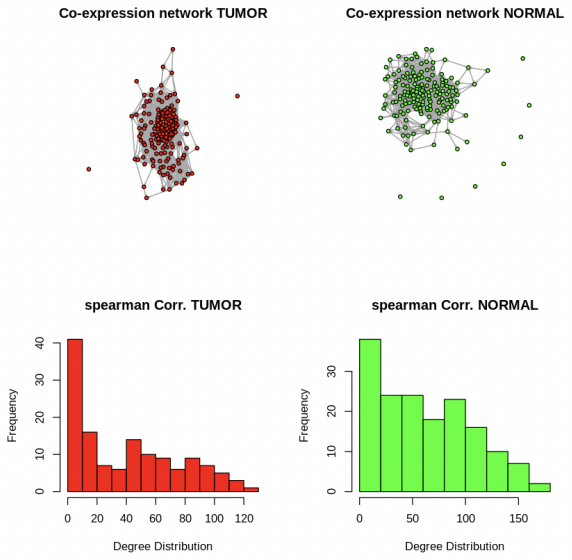
- betweeness: `ENSG00000176928.7`;
- closeness: `ENSG00000176928.7`, `ENSG00000099957.16`, `ENSG00000128266.9`, `ENSG00000103485.19`;
- eigen: `ENSG00000176928.7`, `ENSG00000099957.16`, `ENSG00000128266.9`, `ENSG00000103485.19`, `ENSG00000120885.22`, `ENSG00000152137.8`, `ENSG00000109846.9`.

The results are compared in the following plot.



The normal case has also been analyzed and the relevant plots can be found in the notebook in [GitHub TCGA-PRAD].

B.2 Results obtained using different similarity measures. In the following plots are shown the results obtained applying Spearman correlation (cfr. Figure 2).



References

[AGO] https://atlasgeneticsoncology.org/atlas-explorer?text=

[Bettuzzi et al 2003] S. The new anti-oncogene Clusterin and the molecular profiling of prostate cancer progression and prognosis. Acta Biomed [Internet]. 2003 Aug. 1. Available from: https://www.mattioli1885journals.com/index.php/actabiomedica/article/view/2133

[Brawley 1997] Brawley, O.W. (1997), Prostate carcinoma incidence and patient mortality. Cancer, 80: 1857-1863.

[Cristofani et al 2021] Cristofani R, Piccolella M, Crippa V, Tedesco B, Montagnani Marelli M, Poletti A, Moretti RM. The Role of HSPB8, a Component of the Chaperone-Assisted Selective Autophagy Machinery, in Cancer. Cells. 2021 Feb 5;10(2):335. doi: 10.3390/cells10020335. PMID: 33562660; PMCID: PMC7915307.

[GitHub TCGA-PRAD] https://github.com/leonardoLavagna/TCGA-PRAD

[Kaler et al 2020] Kaler J, Hussain A, Haque A, Naveed H, Patel S. A Comprehensive Review of Pharmaceutical and Surgical Interventions of Prostate Cancer. Cureus. 2020 Nov 22.

[MaryBeth et al 2020] MaryBeth B. Culp, Isabelle Soerjomataram, Jason A. Efstathiou, Freddie Bray, Ahmedin Jemal, Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates, European Urology, Volume 77, Issue 1, 2020

[Potosky et al 1995] Potosky AL, Miller BA, Albertsen PC, Kramer BS. The role of increasing detection in the rising incidence of prostate cancer. JAMA. 1995 Feb 15;273(7):548-52. PMID: 7530782.

[Rawla et al 2019] Rawla P. Epidemiology of Prostate Cancer. World J Oncol. 2019 Apr;10(2):63-89. doi: 10.14740/wjon1191. Epub 2019 Apr 20. PMID: 31068988; PMCID: PMC6497009.

[TCGA-PRAD] https://portal.gdc.cancer.gov/projects/TCGA-PRAD