Finding differentiantly expressed genes in thyroid cancer

Karmele Alapont, Marina Bataller, Nerea Carrón and Judit Garcia

January 31, 2020

1 Abstract

One of the most important changes developed in cancer progression is the alteration of expression patterns inside the cells. This alteration can be caused by many factors. It is interesting to be able to see this changes and an interesting approach is to use RNA-Seq as method to identify and quantify these changes. Here, we performed a RNA-Seq analysis with a thyroid cancer patients dataset. We found 7 genes that were differentially expressed in this cancer, 1 underexpressed and 6 overexpressed. These genes would need further investigation to determine their function in this type of cancer.

2 Introduction

RNA-Seq analysis is a method developed to study expression profiles in different cells. The aim of this analysis is to study the transcriptome, which is the complete set of transcripts of the cell and their quantity, and help us identify the genes that are being transcribed at a specific moment. The expression profile that is obtained from this analysis allows us, for example, to understand better the genes expressed in a specific pathology, such as cancer. Moreover, this is a high-throughput screening (HTS) technology and it is quantitative, which means it provides a measurement of levels of transcripts and their isoforms.

The process used in this analysis is the following. Initially, the RNA is extracted from cells. This is then converted into a library of cDNA fragments, with adaptors attached to the ends of this fragment. After that, each molecule, with or without amplification, is sequenced to obtain short sequences containing 30-400 bp. The last thing before the interpretation of the results is to align the reads to a reference genome or assemble them *de novo*. This way, you obtain a map that shows not only the transcripts present in those cells, but also the level of expression of these transcripts.

We downloaded the RNA-Seq data of thyroid cancer from Recount, an online resource that stores data of RNA-Seq and exon counts of different studies. Among these different studies whose data is stored in Recount there is TCGA, The Cancer Genome Atlas. This project is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) and consists of information about gene expression in 33 different types of cancer. This information is public, it can be downloaded, and it is important for cancer research, in specific fields such as prevention, diagnosis or treatment.

3 Methods

3.1 Packages and tools used

In order to find these overexpressed and underexpressed genes we used some packages from Bioconductor, an open source software for bioinformatics, which provides tools for the analysis and comprehension of high throughput genomic data. The packages we used are:

- ggplot2: data visualization package, used to create graphics.
- BiocManager: tool to install and update Bioconductor packages.

Specifically, the packages used from Bioconductor are:

- SummarizedExperiment: container that contains one or more assays, mainly as matrix-like object numerics.
- DESeq2: differential gene expression analysis based on the negative binomial distribution.
- org.Hs.eg.db: genome wide annotation for Human, based on mapping using Entrez Gene identifiers.
- biomaRt: to retrieve large amount of data uniformly.
- edgeR: analysis of digital gene expression.
- tweeDEseq: to perform RNA-Seq analysis using Poisson-Tweedie distribution.
- GOstats: basic manipulation tools for graphs, hypothesis testing and other simple calculations.

3.2 Data description

We uploaded the RNA-Seq data of thyroid cancer to R with the function load () and the object rse_gene was created. This data is in a Ranged Summarized Experiment (RSE) format, where rows represent ranges of interest and columns represent samples. The object had the information of 58037 genes (rows) analysed in 572 patients (columns). This data was separated into two different groups according to the stage of the patient, early or late. The tumours in stage i and ii were in the group called early and those which were in stage iii and iv were in the other group. In the object rse_stage, a new variable was created to indicated if the tumour was early or late stage.

Figure 1 was created to visualize the distribution of the available data as well as to see if there was any patient whose stage of cancer was not available (NA). Those patients were removed using the following commands:

```
naData <- is.na(rse_gene$GROUP)
rse_gene <- rse_gene[, !naData]
```

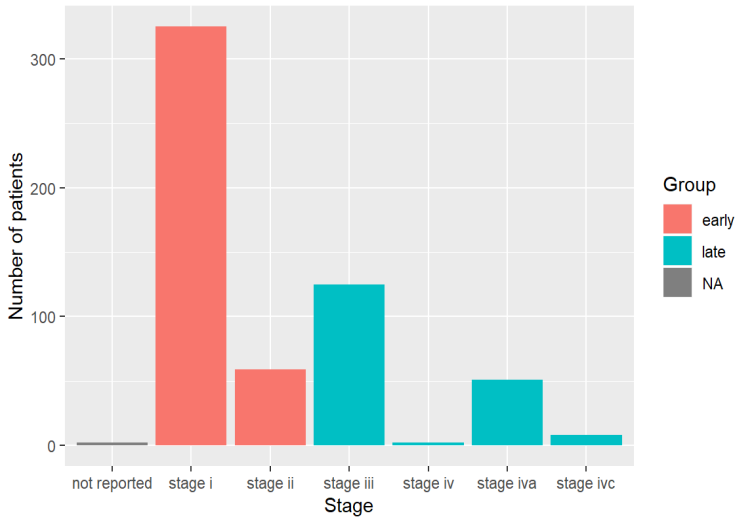

Figure 1: Number of patients in each stage of the tumor. The bars are in blue or pink according to if it is an early or late stage tumor

Finally, we removed 2 patients. 384 tumours were classified as early stage and 186 as late stage, hence, there were two times more data from early stage than late stage.

From the object rse_gene were extracted different information:
• Read counts data using the assay() function and it was available in counts object.
• Phenotype data using the colData() function and it was available in phenotype object.
• Gene information data using the rowData () function and it was available in annotation object. In this object was available the gene id, the length of the gene in bp and the name of the gene for each gene.

We checked whether counts and phenotype objects had the same individuals and they had, hence we still had 570 individuals. We still had the 58037 genes in annotation object, as well.

3.3 Statistic

After the data were prepared, the next step was to normalize the RNA-Seq data. The aim of this step is to remove systematic technical effects that occur in the data to ensure that technical bias has minimal impact on the results (Robinson and Oshlack, 2010).

Even though there are different methods of normalization, we performed only 2:

1. RPKM (Reads Per Kilobase Million). This method corrects for the sequencing depth and the gene length. It was performed with the following formula:

$$RPKM = \frac{\frac{numberofreadsinregion}{regionlengthx10^3}}{totalreadsx10^6}$$

2. TMM (Trimmed Mean of M values). In this method sequencing depth, RNA composition and gene length were considered. This method is recommended to normalized data from different samples and differential expression analyses. It can be performance in R with the function normalizeCounts() of the tweeDEseq Bioconductor package.

We performed MA-plots to check if normalization is needed, figure X. MA-plot represents the log-fold change against the log-average. maPlot () function from edgeR package creates this type of plot. We expected that the majority of genes were not differentially expressed between individuals, the red line (possible trend in the bias related to the mean expression) should be near the 0 of the y-axis. Raw data, Figure 2 , has to be normalized due to the log-fold change decreased over 0 in small values of log-average and in big values, as well. In the range of middle values of x-axis, the log-fold change increased a little bit.

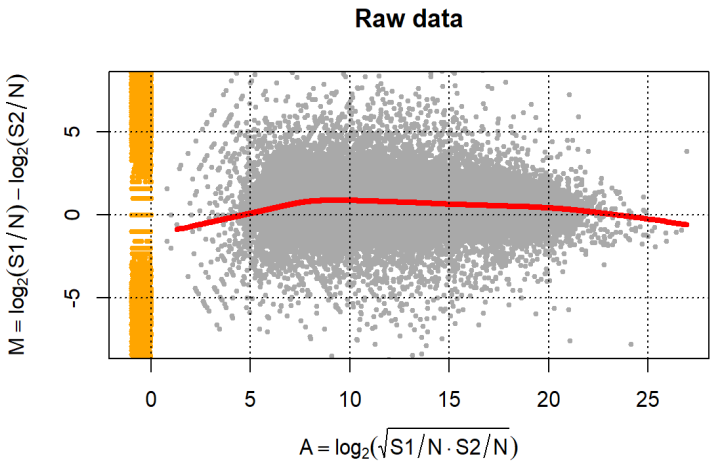

Figure 2: MA-plot with Raw data

We performed both normalization methods with their MA-plots to check whether data improved or not, Figure 3 and 4. In figure 5 we joined the last 3 plots for an easily interpretation. If we compare the Raw data against the RPKM normalization we can not observe any improvement, so in this case, this method of normalization has not helped. While TMM normalization improve the data, due to the red line is more accurately over the M = 0. Hence, in this case, the TMM method is the better option to normalize the data.

4 Results

4.1 Differential expression analysis

Our goal with this analysis was to find differentially expressed genes in tumoral stages and advanced stages in this type of cancer. For this reason, once we had our data prepared we used the R package DESeq2 to test this differential expression.

Since this package has its own normalization method, we used unnormalized data. The following plot shows the results from this analysis:

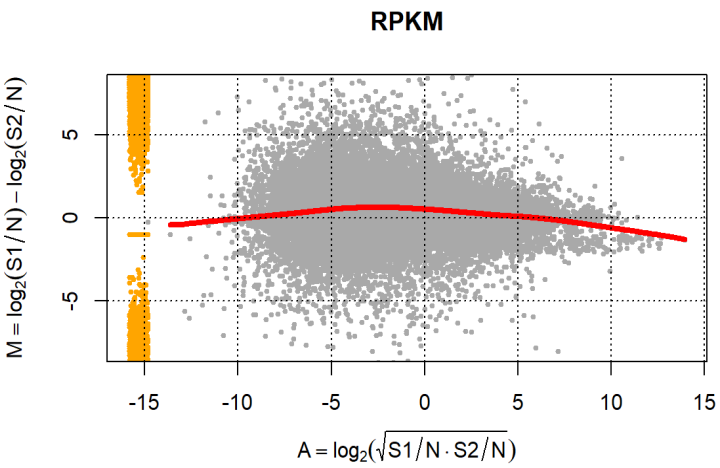

Figure 3: MA-plot with RPKM normalization data

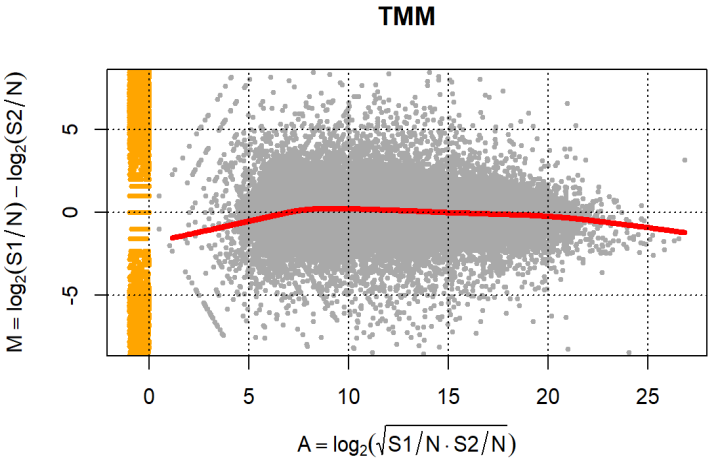

Figure 4: MA-plot with TMM normalization data

This plot (Figure 6) represents the log2 fold-change over the mean of normalized counts for all the samples provided. In red, we have the points that have an adjusted p-value of less than 0.1. The log2 fold-change tells us how many more times a gene is over- or under-expressed in the late stages versus the early stages. A positive value would mean an over-expression and a negative one would mean under-expression. We can see that most points are close to the 0 line, but some get further away from it.

We can filter this result to find the genes that have a significant change of expression in late tumors. We will apply two filters:

• We will keep only the genes whose adjusted p-value is lower than 0.001. After this filter, we keep 2003 genes.

• We will keep only the genes that have a 6 log2 fold-change. We decided to use a slightly lower limit for this filter, since any higher limit we used gave us too few results. We first tried keeping only the genes that have a 10 log2 fold-change, but this only left us with 2 genes. After this lowered filter, we keep 7 genes.

As we can see (Figure 7), 6 of these genes are over-expressed, and one of them is under-expressed.

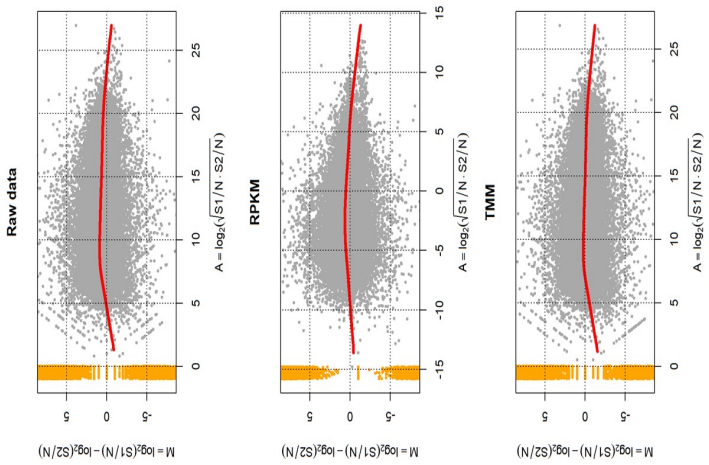

Figure 5: Comparison of the 3 MA-plot obtained from the different data

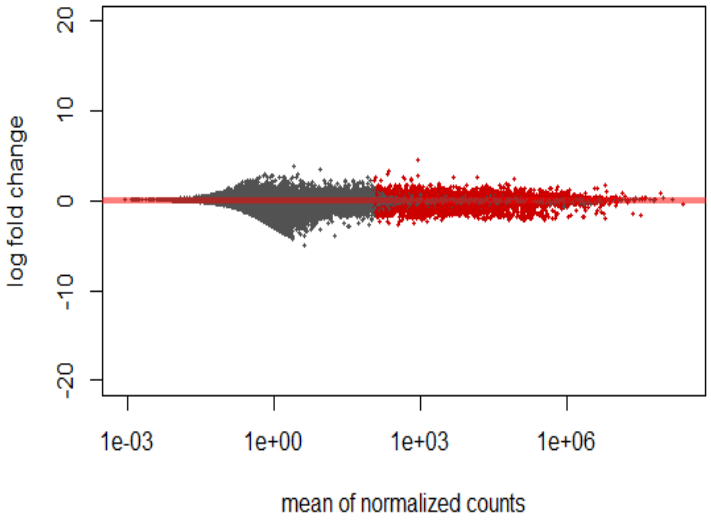

Figure 6: Differentially expressed genes in early vs late thyroid cancer

4.2 Post RNA-Seq analysis: visualization

With these results, we built a volcano plot (Figure 8) for an easier viewing of the data. This is a type of scatterplot that shows the significance in the y axe, and the fold-change in the x axis. It helps us identify quickly changes in large data sets.

Colored in blue we see the genes that have a significant change in expression. On the right of the plot we see the six genes that are over-expressed, and on the left the gene that is under-expressed.

4.3 Post RNA-Seq analysis: enrichment analysis

Once we had a list of genes that have a significant difference in expression, we followed the study with an analysis of the over-expressed genes. We performed a gene set enrichment analysis using the functional annotation of these genes to find more information about them.

We did this analysis using the biomaRt package with the ensembl database and the hsapiens_gene_ensembl as dataset. We tried filtering our list of genes for any NA value, but none was found.

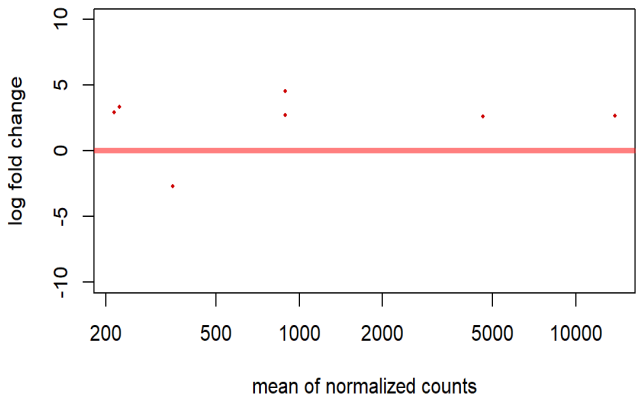

Figure 7: Most differentially expressed genes

Figure 8: Volcano plot

Using the GOstats package we found the Gene Ontology terms associated with these genes, and made a plot with the results (Figure 9). We can see that the GO terms are very scattered: most terms only have one or two genes. The only GO term that all of the genes have is "Multicellular organismal process".

5 Discussion

As we have seen at the results sections there are 6 over-expressed genes and 1 under-expressed. These genes are:

- ENSG00000125414.18 → Gene MYH2
- ENSG00000168530.15 → Gene MYL1
- ENSG00000163092.19 → Gene XIRP2

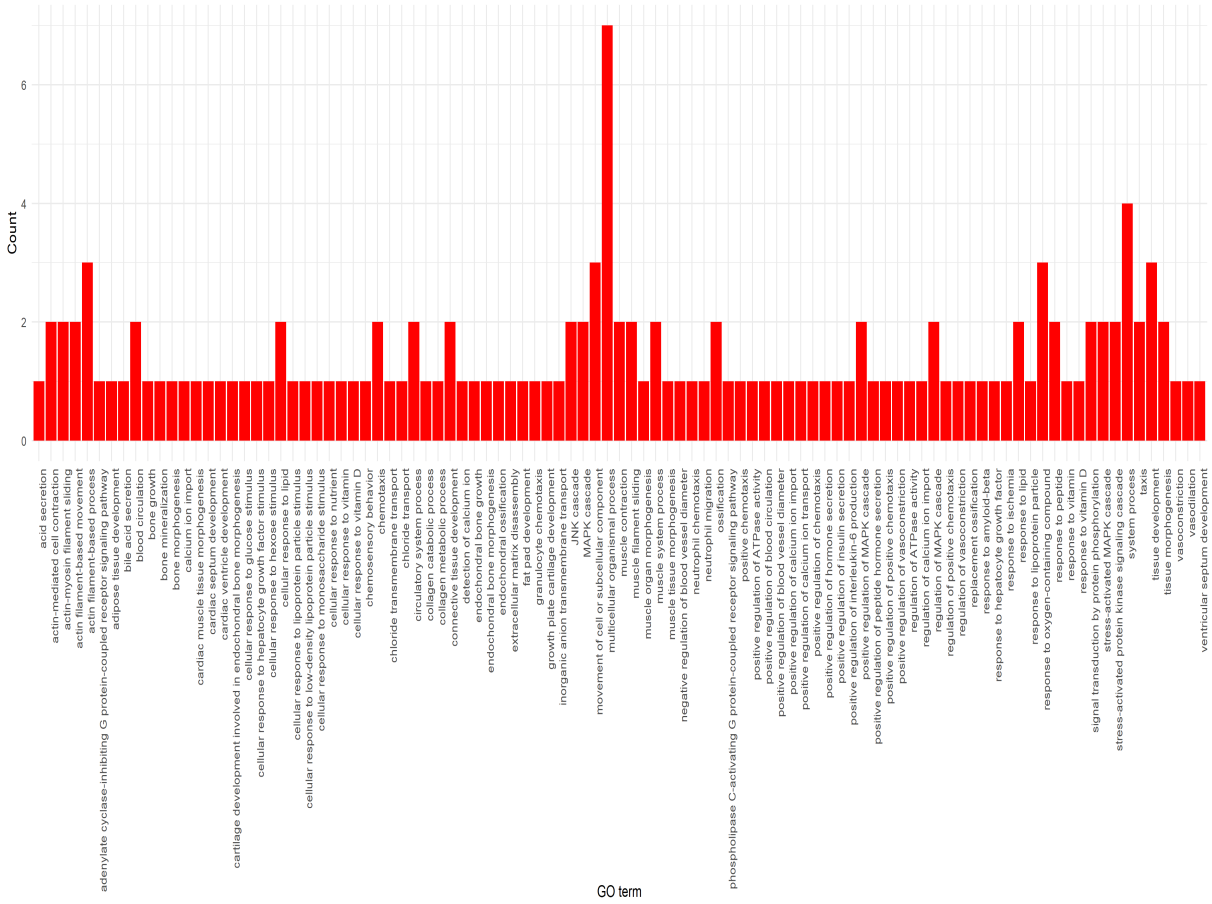

Figure 9: Gene Ontology enrichment analysis

- ENSG00000125571.19 → Gene IL37
- ENSG00000083782.7 → Gene EPYC
- ENSG00000036828.14 → Gene CASR
- ENSG00000137745.11 → Gene MMP13

Now we look if these genes have any implications in thyroid cancer. In order to do so, we looked if this genes encoded proteins and if they did we looked if it had implications in thyroid cancer.

Before starting the analysis we need to take into account that the thyroid is a gland located in the low front of the neck. To see if they had relationship with the thyroid cancer we searched articles that realted them.

5.1 MYH2 gene

It codifies for the myosin heavy chain 2. Myosin is expressed at muscles and its composed by 2 heavy chains and 2 pairs of nonidentical myosin light chains. If we look for the relationship between the thyroid and the gene, there is no bibliography. Looking in The Human Protein Atlas, it says that MYH2 is not prognostic in thyroid cancer, but that is enriched in head and neck cancer.

5.2 MYL1 gene

It codifies for the myosin light chain 1. As mentioned before myosin is expressed at muscles and one of the parts are the myosin light chains. If we look for the relationship between the thyroid and the gene, there

is no bibliography. Looking in The Human Protein Atlas, it says that MYL1 is not prognostic in thyroid cancer.

5.3 XIRP2 gene

It is located in the chromosome 2 and it codifies for the Xin Actin-Binding Repeat-Containing Protein 2. The function of this protein is to protect actin filaments from depolymerization. If we look for the relationship between the thyroid and the gene, there is no bibliography. Looking in The Human Protein Atlas, it says that XIRP2 is not prognostic in thyroid cancer, but that is enriched in head and neck cancer.

5.4 IL37 gene

It is located in the chromosome 2 and it codifies for the interleukin 37. Is a suppressor of innate inflammatory and immune responses involved in curbing excessive information. If we look for the relationship between the thyroid and the gene, there is no bibliography. Looking in The Human Protein Atlas, it says that IL37 is not prognostic in thyroid cancer, but that is enriched in lung cancer.

5.5 EPYC gene

This gene is located in the chromosome 12 and it codifies for the Epiphycan protein. It may have a role in bone formation and also in establishing the ordered structure of cartilage. If we look for the relationship between the thyroid and the gene, there is no bibliography. Looking in The Human Protein Atlas, it says that EPYC is not prognostic in thyroid cancer, but it is in pancreatic cancer (unfavourable) and it is enriched in ovarian cancer.

5.6 CASR gene

This gene is located in the chromosome 3 and it codifies for a calcium sensing receptor. It senses fluctuations in the circulating calcium concentration and modulates the production of parathyroid hormone (PTH) in parathyroid glands (by similarity). If we look for bibliography there are some articles that relate CASR over-expression with thyroid cancer and there is one that looks for a pontential drug using a CASR antagonist (Ding et al.).

5.7 MMP13 gene

This gene is located at the chromosome 11 and encodes a member of the peptidase M10 family of matrix metalloproteinases. This protein plays a role in the degradation of extracellular matrix proteins and it may play a role in cell migration and in tumor cell invasion, among other things. If we look for bibliography there are some articles that relate the MMP13 and thyroid cancer (May et al.).

Looking at the results we have obtained, we concluded that the analysis we have made is good for detecting the genes differentially expressed in thyroid cancer. Nevertheless, we need to take into account that from seven genes that are under or over-expressed we only obtained two that are actually related with thyroid cancer. This could have been because we changed the value of the fold-change threshold in order to have some more genes to analyse, but looking at the ones we obtained before they were MYH2 and MYL1 (both related with myosin) that do not have bibliographical information about their relation with thyroid cancer.

So, we believe that the change that we have made is actually useful for the analysis and that, sometimes, looking for information about a few more genes can make the analysis better.

6 Bibliography

Robinson, M.D., Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11, R25 (2010). https://doi.org/10.1186/gb-2010-11-3-r25.

Wang, Zhong et al. “RNA-Seq: a revolutionary tool for transcriptomics.” Nature reviews. Genetics vol. 10,1 (2009): 57-63. doi:10.1038/nrg2484.

Ding H, Yusof AM et al. ”Localization of CaSR antagonists in CaSR-expressing medullary thyroid cancer.” J Clin Endocrinol Metab. 2013 Nov;98(11):E1722-9. doi: 10.1210/jc.2013-1756

Ma Y, Cang S et al. ”Integrated analysis of transcriptome data revealed MMP3 and MMP13 as critical genes in anaplastic thyroid cancer progression.” J Cell Physiol. 2019 Dec;234(12):22260-22271. doi: 10.1002/jcp.28793