January 2022
Grade in Genetics 2021-22

Adrenal gland cancer: RNA-seq analysis

Universitat Autònoma de Barcelona

Mar Alvarez, Claudia Beneyto, Rocío Calvo and Carla Hijazo

`Professors: Marta Coronado and Nerea Carron`

Current Topics in Bioinformatics

Abstract: *Benign adrenal gland tumors are rare, non-cancerous masses. As part of the endocrine system, the adrenal glands (small, triangular in shape, located up from the kidneys) produce hormones that give instructions to almost every organ and tissue in your body. This type of cancer can manifest at any age. However, it is more likely to affect children under 5 years of age, and adults between 40 and 50 years of age. When adrenal gland cancer is found early, there is scope for care. But if the cancer has spread beyond the adrenal glands, the chances of care decrease. The main objective of this review is the study of adrenal gland cancer by means of RNA-seq analysis. This analysis has been carried out using Rstudio and the packages: ggplot2, ggrepel, BiocManager. The results show that twice as many genes are under expressed as over-expressed in adrenal gland cancer. Those genes of the adrenal gland that respond to stimuli and chemical substances are those that are defferentially expressed in relation to the rest of the categories. Genes for metabolic processes of different chemicals are significant as well.*

Keywords: **Adrenal gland, RNA-Seq Analysis, Adrenal cancer**

Introduction

The adrenal glands are two small triangle-shaped glands located on top of each kidney. Each adrenal gland is about the size of the top part of the thumb. The outer part of the gland is called the cortex and it produces steroid hormones such as cortisol, aldosterone, and hormones that can be changed into testosterone. The inner part of the gland is called the medulla and it produces epinephrine and norepinephrine [1].

When the glands produce more or less hormones than normal, you can become sick. This might happen at birth or later in life. The adrenal glands can be affected by many diseases, such as autoimmune disorders, infections, tumors, and bleeding. Some are permanent and some go away over time. Medicines can also affect the adrenal glands, because they can metabolize some drugs. There are some diseases from others glands can lead to problems with adrenal function, such as the pituitary one's. The pituitary, a small gland at the bottom of the brain, releases a hormone called ACTH that is important in stimulating the adrenal cortex.

Adrenal gland cancer are divide into two types cortex and medulla tumors. Most tumors of the adrenal cortex are benign tumors known as adenomas. These tumors are usually less than 5 centimeters across [2]. They usually occur in only one adrenal gland, but sometimes occurs in both of them. This type of cancer can manifest at any age. However, it is more likely to affect children under 5 years of age, and adults between 40 and 50 years of age. When adrenal gland cancer is found early, there is scope for care. But if the cancer has spread beyond the adrenal glands, the chances of care decrease [3].

A RNA-seq was perform in this review because is a recent approach to carry out expression profiling using high-throughput sequencing (HTS) technologies. In the past decade, microarrays were predominantly used for this kind of work, but since price of sequencing has been reduce, RNA-seq has become the preferred option to simultaneously measure the expression of tens of thousands of genes for multiple samples. In this review we walk through a gene-level RNA-seq differential expression analysis using Bioconductor packages to find genes over- or under-expressed in adrenal gland cancer patients.

Methods

Packages and tools used

To perform the data analysis, *Rstudio* was used, it is an integrated development environment for the R programming language, dedicated to statistical computing and graphics. The installation of the *ggplot2*, *ggrepel* and *BiocManager* packages is precise, all of them were provided by the *Bioconductor team. Bioconductor* has several packages that allow us to import and process raw sequencing data, and upload gene annotations. This enables high-throughput sequence data analysis, including RNA-seq. Tools that were used:

- *SummarizedExperiment*
- *DESeq2*
- *org.Hs.eg.db*
- *biomaRt*
- *edgeR*
- *tweeDEseq*
- *GOstats*
- *tweeDEseqCountData*
- *annotate*

Data description

Experimental data was extracted from The Cancer Genome Atlas (TCGA). TCGA is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, comprising more than two petabytes of genomic data, has been made publicly available, and this genomic information helps the cancer research community to improve the prevention, diagnosis, and treatment of cancer.

Recount is an online resource consisting of RNA-seq gene and exon counts for different studies, including TCGA data. The RNA-seq data of Adrenal gland cancer was extracted from there. The data set is available at this link `http://duffel.rail.bio/recount/v2/TCGA/rse_gene_adrenal_gland.Rdata` and can be downloaded.

Statistical tests

Normalization

It is necessary to normalize RNA-seq data for several reasons:

- The number of counts is related to sequencing depth:

  Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in Sample A relative to Sample B, however this is a consequence of Sample A having double the sequencing depth.

- The number of counts is related to transcript length:

  Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, Gene X and Gene Y have similar levels of expression, but the number of reads mapped to Gene X would be many more than the number mapped to Gene Y because Gene X is longer.

- The number of counts is proportional to the mRNA expression level:

  A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples, and is particularly important when performing differential expression analyses.

This theoretical explanation was drawn from this GitHub [4].

The aim of normalization is to remove systematic technical effects that occur in the data to ensure that technical bias has minimal impact on the results. [5] There are different methods to normalize the counts [Evans et al., 2017]:

- **RPKM:** [Mortazavi et al., 2008] Reads Per Kilobase Million. This method corrects for the sequencing depth and the gene length. It's a non-sophisticated normalization method. Counts are divided by the transcript length (kb) times the total number of millions of mapped reads:

$$RPKM = \frac{\frac{number\ of\ reads\ in\ region}{region\ length \times 10^3}}{total\ reads \times 10^6}$$

- **TMM:** [Robinson and Oshlack, 2010] Trimmed Mean of M values. This method was developed to address this issue: The proportion of reads attributed to a given gene in a library depends on the expression properties of the whole sample rather than just the expression level of that gene. Therefore, the method accounts for sequencing depth, RNA composition, and gene length.

Both methods allows comparisons between genes within a sample, but only TMM is recommended for between sample comparisons and differential expression analyses.

The MA-plots are used to check whether normalization is needed or not. There is a function in the *edgeR* package designed to create such a plot, the *maPlot()* function.

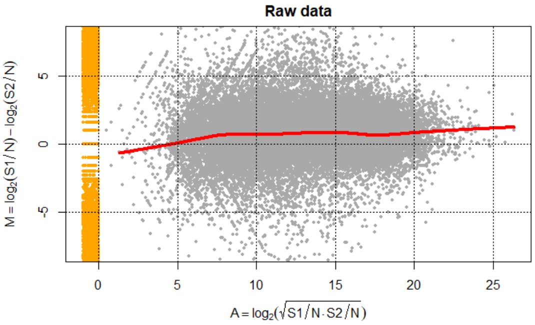

Figure 1: *Raw data MA-plot. Graphic made with Rstudio.*

This plot represents the log-fold change (M-values, i.e. the log of the ratio of level counts for each gene between two samples) against the log-average (A-values, i.e. the average level counts for each gene across the two samples). From a MA-plot one can see if normalization is needed or not.

One expects that the vast majority of genes are not differentially expressed between individual, thus having a symmetrical distribution on the plot with most of the genes at the 0 (y = 0). A lowess fit (red line) is plotted underlying a possible trend in the bias related to the mean expression. As seen in the graph 1, it is necessary to normalize the data because they do not follow a normal. There's an upward bias on the graph.

Results with figures

We extracted the data collected in the RSE data and explored the data using the dim() function, we found that there are 58037 genes analyzed in a cohort of 266 patients with adrenal gland cancer. We then divided the patients into groups according to the tumor stage in which they had been diagnosed. The groups are collected in the following Figure 2.

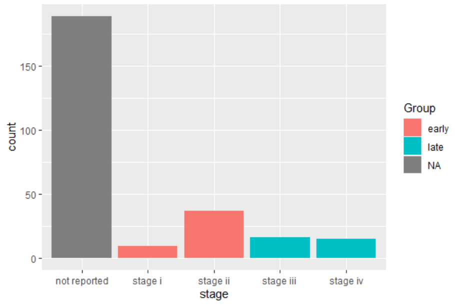

Figure 2: *Raw data. Graphic made with Rstudio.*

As it shown, there were many patients whose stage of cancer was not recorded (189). For this reason we only kept 77 that had all the information accessible and complete, 46 of them were diagnosed in early stage, whereas 31 were in late stage. It was checked that all of the patients in this cohort meets both requirements: being in the counts dataset and in the phenotype dataset. Finally, we have the information of 58037 genes with the gene annotation provide of gene ID, length in bp and symbol.

The normalization of this dataset was made with the aim to remove systematic technical effects that occur in the data to ensure that technical bias has minimal impact on the results. Two different methods to normalize the count were ran: RPKM and TMM.

- RPKM normalization allows normalizing read counts simply applying the previous formula shown in Methods.
- TMM normalization method is implemented easier because it is in the *tweeDEseq* Bioconductor package, inside the *normalizeCounts()* function, as we say in Methods.

Below it shown a *MA-plot* of the data with the RPKM and TMM normalization shown in Figure 3.

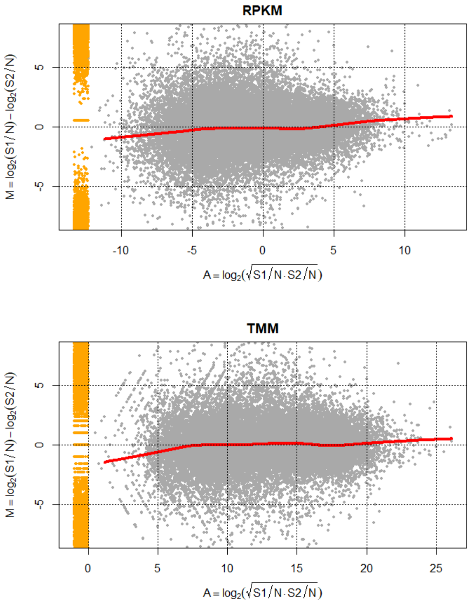

Figure 3: *RPKM and TMM normalization MA-plot. Graphic made with Rstudio.*

With no doubt the TMM normalization achieve better results in comparison with the raw data and the RPKM normalization.

Differential expression analysis

For the differential expression analysis the R package *DESeq2* was used. This package allows researchers to test differential gene expression analysis based on the negative binomial distribution as *Love, et al. (2014)* showed in their article.

The starting point of a *DESeq2* analysis is a count matrix with one row for each gene and one column for each sample. The data is unnormalized reads count, because this package has its own normalization method. This function requires the *SummarizedExperiment* object and the design, which in this case is given in the variable *GROUP* because we want to compare early vs. late tumor stages.

When the variable *GROUP* was fixed, plot with the *plotMA* function is now possible. This function allows the graphical representation of the log_2 fold-change over the mean of normalized counts for all the samples in the *DESeqDataSet*. It can be seen in Figure 4 where the points colored in blue have p-values lower than 0.1 in the adjustment and the points which fall out of the window are plotted as open triangles pointing either up or down.

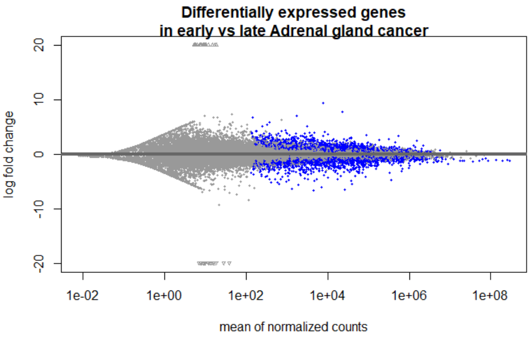

Figure 4: *RPKM MA-plot. Graphic made with Rstudio.*

With this analysis we were able to find the genes that are significantly overexpressed or underexpressed in late tumours. The following filters were apply for it:

- 329 genes were kept because their adjusted p-value is lower than 0.001.
- 45 genes were kept because they have a 10 log_2 fold-change. This criteria is more stringent because we reduce in 7 times the number of genes that we kept.

A MA-plot was used again for show the results of most differentially expressed genes, see in Figure 5.

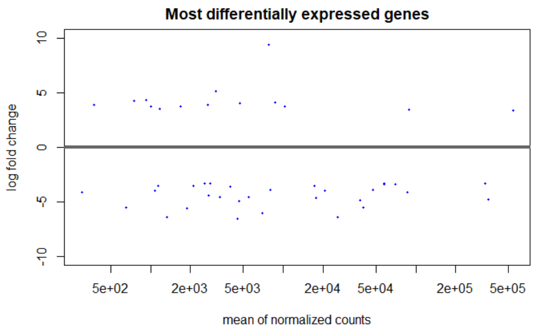

Figure 5: *Most differentially expressed genes MA-plot. Graphic made with Rstudio.*

When we represent only the differentially expressed genes we have twice as many genes underexpressed as overexpressed, 29 versus 14. In the MA-plot are missing two genes of the counting of the most differentially expressed genes.

- Underexpressed: 29
- Not differentially expressed: 54561
- Overexpressed: 14

RNA-Seq analysis

A volcano plot was used to identify changes in large data set for plot significance vs. fold-change on the y and x axes, respectively. This plot is essentially an scatter plot, constructed by plotting the negative logarithm of the P-value on the y-axis (usually base 10). This results in data points with low P-values (highly significant) appearing toward the top of the plot. The x-axis is the logarithm of the fold change between the two conditions (usually base 2). Each point (gene) will be colored based on the filtering (*filter*). Below in Figure 6 its shown a volcano plot that have annotated the most significant (padj < 0.0001) genes.

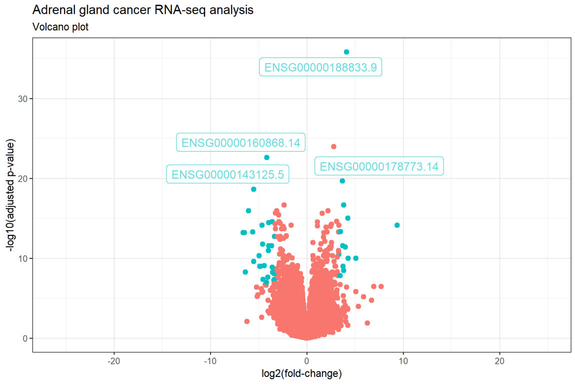

Figure 6: *Volcano plot DE genes. Graphic made with Rstudio.*

Finally, the gene expression data was interpreted through a gene set enrichment analysis based on the functional annotation of the differentially expressed genes. This is useful for finding out if the differentially expressed genes are associated with a certain biological process or molecular function.

NA values were cleaned and only those Entrez gene ids for which we have the information were kept a total of 41. Go enrichment analysis was perform for the differentially expressed genes. In the GO analysis is shown in Figure 7 information for a total of 20688 human genes is collected.

As discussed in the introduction, the adrenal gland must respond to stimuli and chemicals in order to produce steroid hormones on demand from the body. Additionally, the production of neurotransmitters such as adrenaline and noradrenaline is also made in the adrenal gland. Other processes that we see that are highly significant are the metabolic processes of different chemicals. These results fit biologically with the organ that was studied in the RNA-seq analysis and therefore, they are valid and robust results.

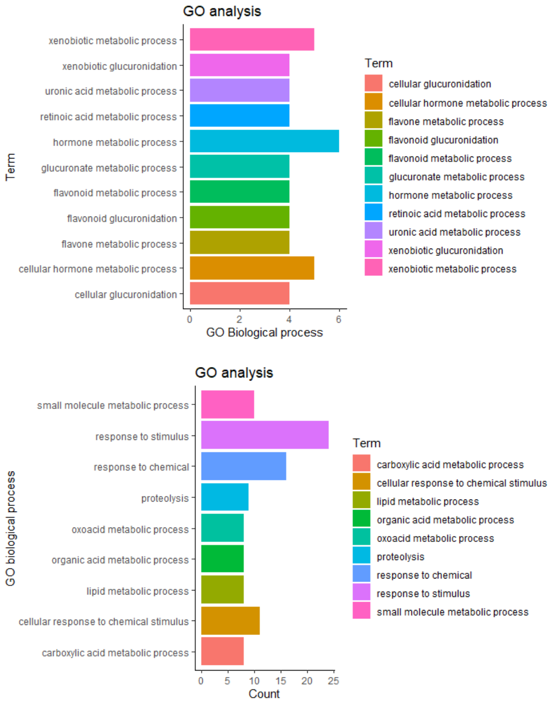

Figure 7: *Go analysis. Graphic made with Rstudio.*

Discussion

The analysis identified that the main cause of adrenal gland cancer is the underexpression of genes, much more than their overexpression. Thanks to GO analysis, show in figure 7, we know that about half of the differentially expressed genes regulate or code for proteins involved in cellular responses to chemical stimuli. In this organ it is especially sensitive to this type of stimuli because it is a hormone-producing organ. Other biological processes affected are proteolysis, carboxylic acid, oxoacid, small molecule, lipid, and organic acid metabolic processes. This is critical considering that they are the main precursors needed to generate hormones. As was said the adrenal cortex makes 3 main hormones: cortisol, aldosterone, and dehydroepiandrosterone (DHEA) that carefully control metabolism, blood pressure, and body features, such as hair growth and body shape. Furthermore, the adrenal medulla makes 3 more hormones: epinephrine, norepinephrine, and dopamine. These catecholamines control the body's responses to stress, including the "fight or flight" adrenaline surge. If any of them falls out of control or are poorly synthesized, it can mean the loss of homeostasis in the organism.

Among the affected metabolic processes we highlight:

- xenobiotic metabolic process
- uronic acid metabolic process
- retinoic acid metabolic process
- glucoronate metabolic process
- flavonoid metabolic process
- flavone metabolic process
- cellular hormone metabolic process

Cellular glucuronidation is also affected, especially flavonoid and xenobiotic. The glucoronidation is involved in drug metabolism, it occurs mostly in the liver, but it is also found in the adrenal gland.

Genes annotated in Figure 6 code for:

- *ENTPD8*(ENSG00000188833.9) that codes for ectonucleoside triphosphate diphosphohydrolase 8.

- *CYP3A4* (ENSG00000160868.14) that codes for cytochrome P450 family 3 subfamily A member 4.

- *PROK1* (ENSG00000143125.5) that codes for prokineticin 1 a receptor ligand.

- *CPNE7* (ENSG00000178773.14) that codes for Copine-7 a calcium-dependent phospholipid-binding protein.

Conclusion

The most differentially expressed genes in adrenal gland cancer play a important role in metabolism process and in cellular response. In this review, just a little sample of all the genes behind adrenal gland cancer were discussed. RNA-seq analysis may be decisive in the future for research community to improve the prevention, diagnosis, and treatment of many types of cancer.

References

[1] Cáncer de la glándula suprarrenal - Síntomas y causas - Mayo Clinic. (21 of January of 2021). https://www.mayoclinic.org/es-es/diseases-conditions/adrenal-cancer/symptoms-causes/syc-2035102

[2] Glándulas suprarrenales: MedlinePlus enciclopedia médica. (21 of January of 2021). https://medlineplus.gov/spanish/ency/article/002219.htm

[3] Adrenal Gland Cancers: Symptoms, Diagnosis & Treatment - Urology Care Foundation. (21 of January of 2021). https://www.urologyhealth.org/urology-a-z/a_/adrenal-gland-cancers.

[4] Normalization information retrieved from [Introduction to DGE]. https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html.

[5] A scaling normalization method for differential expression analysis of RNA-seq data *Robinson, Mark D., Oshlack, Alicia Genome Biology, (2010), 1-9, 11(3).* https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25