Analysis of a TCGA RNA-seq data set on lung squamous cell carcinoma

Carbonell Gamón, Gabriel[*,1], **Matabacas Enebral, Aleix**[*,1] **and McKitrick, Austin**[*,1]
[*]Master Programme on Bioinformatics for Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Lung cancer, including lung squamous cell carcinoma, is the most lethal type of cancer and lacks early diagnostic tools and treatment options. Research to understand the genes and biological pathways involved in this cancer subtype could provide information to solve this problem. The aim of this study was to analyze RNA expression profiles from patients with this specific cancer subtype in order to find differently expressed genes and pathways that are relevant for this disease. To do this, a differential expression analysis and Gene Ontology analysis were conducted on transcriptomic data using a lung squamous cell carcinoma dataset from The Cancer Genome Atlas project, in which healthy and tumor tissue samples from the same patients were compared. The results of this study showed that 1451 genes were significantly differentially expressed. The subsequent functional analysis suggested that those differentially expressed genes found between cancer and non-cancer samples were heavily involved in signal transduction and cell metabolism. These genes and pathways could be useful starting points to accelerate diagnostics and improve treatments for lung squamous cell carcinoma.

KEYWORDS lung cancer; non-small cell lung cancer; transcriptomics; squamous cells; differential expression, gene ontology, TCGA

Introduction

Among all cancer types, lung cancer is the most lethal among both males and females , being responsible for 28% of cancer-related deaths (Stewart *et al.* 2014). The life expectancy for those who have lung cancer is very poor and usually is not diagnosed until the cancer's advanced phase. In addition to a late diagnosis, research has shown that patients only have a 16% survival rate after 5 years of their diagnosis (National Cancer Institute 2019b).

One of the most common types of lung cancer is lung squamous cell carcinoma (LUSC), which typically occurs in the larger bronchi of the lung. Although smokers have a high prevalance of LUSC, non-smokers can suffer this disease too (Park *et al.* 2017). This is a non-small cell lung cancer, which has been found to respond poorly to radiotherapy and chemotherapy in comparison to other kinds of cancer (National Cancer Institute 2019a). The most common treatment is surgery, which is highly invasive to the patient and doesn't always cure the cancer. An early diagnosis of this cancer could allow for early treatment making the need for surgery avoidable.

Considering these factors, research to characterize the genes and pathways related to LUSC and to increase the prediction and diagnosis capacity is necessary. Previous studies on this topic show the existence of biomarkers that may be used for diagnostic and prognostic purposes, as well as molecular targets for the treatment of this group of cancer (Xiao *et al.* 2018; Hayes *et al.* 2014). Unfortunately, this study lacked sensibility in the detection of sub-types of cancer, as it targeted all non-small lung cell carcinomas.

Therefore, the present study aimed to contribute to the characterization of lung squamous cell carcinoma. Using data profiled from The Cancer Genome Atlas (TCGA), RNA expression profiles were analyzed from patients with LUSC in order to find differently expressed genes and pathways that are relevant for this disease, setting the basis to develop diagnostics and improve treatments.

Materials and Methods

Data Collection

Data was obtained from The Cancer Genome Atlas program (TCGA) (Weinstein *et al.* 2013). The data consisted of a table

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Sunday 16th June, 2019
[1]gabriel.carbonell01@estudiant.upf.edu, aleix.matabacas01@estudiant.upf.edu, austin.mckitrick01@estudiant.upf.edu

of raw counts containing 20115 genes and 553 samples, organized as a *SummarizedExperiment* object. Of those samples, 502 belonged to tumor tissue and 51 to normal tissue.

Statistical Analysis Design

Data analyses were performed using the R programming language, supported by open source packages from the Bioconductor project. Specific versions used for this study can be found under the *Session information* (Morgan *et al.* 2018b) sections from the *Supplementary Materials*. The data was read using the *SummarizedExperiment* package, and most plots were created with *geneplotter* (Gentleman and Biocore 2018). The analysis focused on studying the differences between paired normal and tumor samples. For this, the data set was subsetted by filtering for those samples whose patient code was the same.

Quality Assessment and Normalization

After filtering, data quality was assessed and normalization procedures were applied. This required the transformation of the *SummarizedExperiment* object into a *DGElist* object by means of the *edgeR* package (Robinson *et al.* 2010; McCarthy *et al.* 2012). Changes were kept in sync for both object types. Within-sample normalization was achieved by calculating the counts per million (CPM), which were then log2 transformed. Next, the distributions of library sizes and expression levels among samples and expression levels among genes were examined visually with plots. Based on the later one, genes below 1 logCPM were considered lowly-expressed and were filtered out. Between-sample normalization was done by calculating normalization factors with *edgeR*, and the results of this step were assessed with MA plots. Possible batch effect sources were analyzed by examining the distribution of the elements of the TGCA barcode across samples as surrogate variables. Elements for which a suspicious distribution was found were analyzed by means of hierarchical clustering and multidimensional scaling plot. This visualization was done by removing tumor samples from portion analytes which did not have any normal samples avoiding zeros in the cross-reference table of the outcome of interest and the portion analyte. Normal pairs of these samples were also removed from other portion analytes for the visualization. Samples that did not meet the quality criteria were discarded.

Differential Expression Analysis

The *limma-voom* pipeline from the *limma* package (Ritchie *et al.* 2015) was used to perform differential expression analysis. A linear model consisting of the type of sample as the outcome of interest and adjusting for the paired design of the analysis by using the patient ID were created. No further adjustments for any known effect were done to the model. The *sva* package (Leek *et al.* 2019) was used to estimate surrogate variables as indicators of unknown variability, which were then added to the model. The performance of statistical tests was assessed with their distributions of raw p-values and moderated t-statistics. After the statistical test, a false discovery rate (FDR) of 1% was used for type I error correction, and only those genes below that threshold and with a low fold change higher than two or lower than minus two were considered differentially expressed. The differential expression analysis results were diagnosed with volcano and MA plots built using *limma*.

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was completed using the *GSEABase* package (Morgan *et al.* 2018a) and the C2 collection from the Broad Institute's *MSigDB* gene set collections was used for the analysis. The collection was imported using the *GSVAdata* package (Castelo 2018), and contained 29340 genes and 3272 curated computational gene sets. The analysis calculates an enrichment score for each gene set, which provides information on the changes in gene expression by individual genes in the gene set.

After the individual identifiers in the C2 gene sets and the data set being analyzed were properly mapped to one another, an incidence matrix was created. This matrix indicated what individual genes belonged to the gene sets and the data set being analyzed. A *1* denoted that the gene was a member of that data set, while a *0* denoted that the gene was not a part of that data set. After this, all genes that were not part of data set, or had a *0*, were discarded.

SVA was then conducted with the same procedure outlined previously. The generated t-statistics were stored in a vector structure and the z-scores were calculated to identify shifts in gene expression within a gene set. A filter was completed to remove genes sets containing less than 5 genes. Finally, a one sample z-test was calculated and a conservative multiple testing adjustment was administered to find if any gene sets were candidates for differentially expressed genes.

Gene Set Variation Analysis

In addition to the GSEA, a gene set variation analysis (GSVA) was performed with the package *GSVA* (Hänzelmann *et al.* 2013). This differed from the GSEA previously administered by utilizing a gene set by sample matrix rather than a gene by sample matrix. This matrix was also created with the number of gene sets containing five or more genes and an enrichment score that indicated sample-wise gene-level summaries of expression. The enrichment scores were then used for gene expression values in the differential expression analysis. The genes that were differentially expressed at 1% FDR were then identified.

Gene Ontology Analysis

Using the differentially expressed genes identified in the differential expression analysis, a Gene Ontology analysis was performed with the package *GOstats* (Falcon and Gentleman 2007). A parameter object was built that specified the gene universe of interest, the set of differentially expressed genes, the ontology, and other information. The ontology selected was *BP*, which matched genes to the Biological Processes associated with the GO Terms. A conditional test was completed and the output was filtered by gene sets containing between 5 and 250 genes. In addition, GO Terms with an Odds Ratio of 1.5 or greater.

Data Availability

The transcript reads used in this analysis can be obtained from the The Genome Cancer Atlas project website. The code used for all the analysis is provided in the supplementary materials, as well as the resulting files of the different filtering steps and analyses under the *results* directory.

Results and Discussion

Quality Assessment

The data set consisted of 20115 genes, 51 tumor samples and 502 normal samples; but after filtering for those samples which were paired, the number was reduced to 102 samples, maintaining the same number of genes. Two normalization steps were needed

to analyze the data: within-sample normalization, to compare across features in a sample, and between-sample normalization to compare a feature across samples. Within-sample normalization was log-transformed to improve the distributional properties. The library size distribution, displayed in the *Figure 1* from the Supplementary Materials, showed a not uniform sequencing depth in which tumor and normal samples were randomly distributed. Gene expression across samples was explored with MDS plots (Supplementary Materials *Figure 2*) for tumor and normal samples separately, without showing any significant problem.

The distribution of expression levels among genes (Supplementary materials *Figure 3*) motivated to set a minimum threshold of expression of 1 log CPM to prevent downstream artifacts in further analyses. After filtering out those genes below that threshold, 11872 genes and 102 samples remained.

After normalization of the data, expression-level dependent biases were checked using MA plots (Supplementary Materials, Figures 4 and 5), which were not observed in either normal or tumor samples in significant levels. The tumor sample from patient *8623* showed in the plot that it was the most different from the average of the others.

Although the portion of analyte stood out as a potential batch effect source, the hierarchical clustering performed with 78 samples showed that those were clustering primarily by type, with the different portions present in both clusters (Supplementary Materials, Figure 6). This reduction of the number of samples was done in order to avoid having zeros in the cross-classification of the outcome of interest with the portion analyte. The resulting plots led to the consideration that this variable was not inducing batch effect. In any case, it was observed that the tumor sample from patient *8623* clustered along the normal samples. The sample was then discarded based on this observation and its MA plot, and its normal pair was also discarded to keep the paired design, leaving a total of 100 samples and 11872 genes. Errors in the annotation of this sample could be the reason for this behaviour, but as there was no access to data prior to annotation, no more insight on the reasons could be provided.

Differential Expression Analysis

The *sva* package estimated 12 surrogate variables, which represent unknown sources of variation. The number of variables could be reduced by fine tuning the model, requiring more analysis of the phenotypical variables present in the data. After including the variables into the model, the *limma-voom* pipeline estimated 9137 significantly expressed genes under a FDR cutoff of 1%. Of this genes, only those which logarithmic fold change was over two or below minus two were considered for further analysis, leaving 1451 genes being called differentially expressed, of which 896 were down-regulated and 555 up-regulated.

Performance of the statistical test done by means of the distribution of raw p-values, shown at Figure 1, and a QQ plot of the moderated t-statistics (Supplementary Materials, Figure 11) showed that diagnoses where correct. The distribution of raw p-values was uniform except for a peak at low p-values, which corresponded to the differentially expressed genes, while the QQ plot showed that most genes were far from the diagonal, showing the significance and amount of the discovered differentially expressed genes.

Regarding the chromosome distribution of those differentially expressed genes passing both the 1% FDR and the log fold change cutoffs, it did not correspond to a distribution based on the chromosome size. This also acts as an indicator of the quality of the results, as a distribution ordered by size is not expected. Chromosomes with more genes related to processes that are biologically relevant for the context of cancer are expected to rank higher. In this case, chromosomes 1, 19 and 11 were on the top of the ranking (Supplementary Materials, Figure 10).

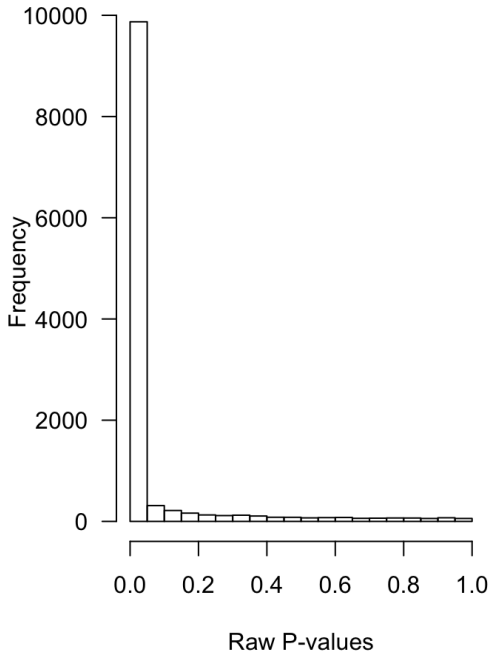

Figure 1 Raw p-values for the differential expression analysis results. The figure shows a uniform distribution for not differential expressed genes, which is expected under the null hypothesis, while there is a peak at low p-values for the truly DE genes.

Finally, the results were diagnosed by means of a volcano plot, shown in Figure 2. It displays the significance of each gene against its level of differential expression, and no unexpected trends are found.

Gene Set Enrichment Analysis

Functional enrichment analysis techniques, such as a GSEA, are valuable tools since they have the ability to detect sensitive and small changes in differential expression at the gene set level. Although 1451 differentially expressed genes were obtained from the differential expression analysis, tests of functional enrichment were explored to compare differences among techniques.

After mapping the C2 data and the data being analyzed, 9963 genes among 100 samples were left for the SVA portion of the GSEA. After conducting a SVA and filtering out gene sets with less than 5 genes, 9963 genes among 2760 gene sets were left from the original 3272 of the C2 collection. A one sample z-test and a conservative multiple testing adjustment was conducted and 2313 gene sets were showed to be differentially expressed.

Gene Set Variation Analysis

In addition, a GSVA was completed as part of the functional enrichment analysis. This differed from the GSEA previously administered by utilizing a gene set by sample matrix rather than a gene by sample matrix, which allows for the pathway

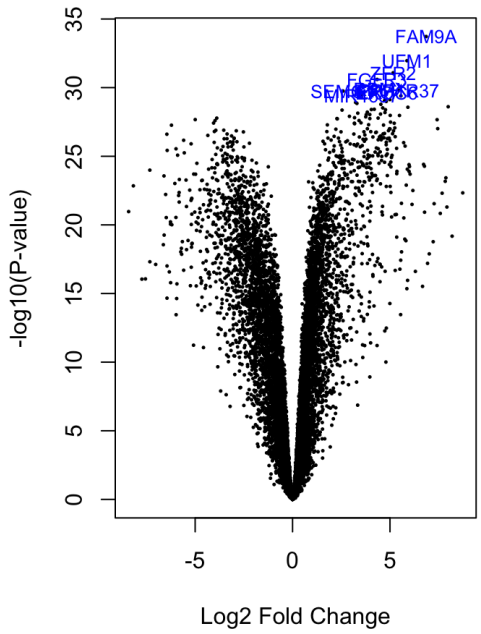

Figure 2 Volcano plot of the differential expression analysis results. Base 10 log-transformed *p-values* are shown in the *y* axis as a measure of significance, while the base 2 logarithm of the fold change is shown in the *x* axis as a measure of the differential expression level. The top 10 differentially expressed genes are labeled. Labels are: FAM9A, UFM1, ZFR2, FGFR3, RBMX, SEMG1, PPP1R37, ERH, EXOC6, MIR4687.

enrichment for each individual sample to be analyzed. After the data was normalized and properly filtered, the GSVA found 1877 gene sets to be differentially expressed at 1% FDR.

Techniques such as GSEA and GVSA are often used only when differentially expressed genes are not found in the DE analysis. Since this study identified 1451 differentially expressed genes in the DE analysis, these genes were selected to complete the GO analysis. Results on GSVA and GSEA are given as preliminary, as the proper analysis of the enriched sets was not performed yet. A possible future direction could explore the GO analysis using the gene sets found in the GSEA and GVSA from this study.

Gene Ontology Analysis

To extract knowledge from the RNA expression profiles, the usage of algorithms to detect DE genes is not enough (Dudoit *et al.* 2002). The biological function of the detected genes should be studied in order to know what functions and pathways are altered. In this sense, the Gene Ontology Consortium homogenizes the format and vocabulary of the biological data obtained from research to make it reusable(Ashburner *et al.* 2000). This allowed for the use of the obtained differentially expressed genes as a query in its database, reducing the amount of experimental studies and accelerating research.

After performing a conditional test and the appropriate filtering, 33 GO Terms were identified. The specific differentially expressed genes associated with each GO Term were identified as well. The results of the GO analysis suggested that the differentially expressed genes discovered in this study were highly involved in signal transduction. Cancer samples had differentially expressed genes associated with signal complex assembly (OR = 27.78, P = 1.14x10-3), the collagen-activated tyrosine kinase receptor signaling pathway (OR = 9.26, P = 6.44x10-3), and adenylate cyclase-activating G protein-coupled receptor signaling pathway (OR = 2.64, P = 5.21x10-4) in comparison to normal samples. Past research using GO analysis has found the parent GO term associated with signal transduction to be significant among the differentially expressed genes studied (Long *et al.* 2019). However, this study didn't find the specific child-class GO terms found in the present study. Perhaps the genes controlling for these specific signaling pathways could be helpful in the development of diagnostics and treatment for this kind of lung cancer Table 1 summarizes the highlighed results from the Gene Ontology analysis.

In addition to signal transduction, the differentially expressed genes were also involved in cell activity, mainly cell metabolism. The GO analysis showed cancer samples had differentially expressed genes associated with the negative regulation of mRNA splicing (OR = 4.63, P = 7.26x10-3), and mRNA splice site selection (OR = 4.28, P = 2.81x10-3). Splicing plays an important role in essential biological processes, including cancer, and past research suggests a relationship exists between the dysregulation of mRNA splicing and lung cancer (Pio and Montuenga 2009). Therapeutic approaches that act on the regulation of mRNA splicing could be a helpful in treating lung cancer.

Concluding Remarks

In conclusion, this study successfully identified differentially expressed genes and associated biological pathways unique to LUSC samples in this data. Among those pathways, it appears that signal transduction and cell metabolism were highly enriched. Therefore, the genes involved in these pathways could be important to developing diagnostics and treatments for lung squamous cell carcinoma.

Acknowledgements

We would like to thank our instructor Robert Castelo and our classmates for all the help and ideas that they have given us during this project. Also thanks to Eric Climent, our friendly statistician, for helping us with its past experience with RNA-seq experiment results.

Literature Cited

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, *et al.*, 2000 Gene ontology: tool for the unification of biology. Nature genetics **25**: 25.

Castelo, R., 2018 *GSVAdata: Data employed in the vignette of the GSVA package*. R package version 1.18.0.

Dudoit, S., Y. H. Yang, M. J. Callow, and T. P. Speed, 2002 Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Statistica sinica pp. 111–139.

Falcon, S. and R. Gentleman, 2007 Using GOstats to test gene lists for GO term association. Bioinformatics **23**: 257–8.

Gentleman, R. and Biocore, 2018 *geneplotter: Graphics related functions for Bioconductor*. R package version 1.60.0.

Hänzelmann, S., R. Castelo, and J. Guinney, 2013 GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics **14**: 7.

Hayes, J., P. P. Peruzzi, and S. Lawler, 2014 Micrornas in cancer: biomarkers, functions and therapy. Trends in molecular medicine **20**: 460–469.

Leek, J. T., W. E. Johnson, H. S. Parker, E. J. Fertig, A. E. Jaffe, *et al.*, 2019 *sva: Surrogate Variable Analysis*. R package version 3.30.1.

Long, T., Z. Liu, X. Zhou, S. Yu, H. Tian, *et al.*, 2019 Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. Molecular medicine reports **19**: 2029–2040.

McCarthy, D. J., Chen, Yunshun, Smyth, *et al.*, 2012 Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic Acids Research **40**: 4288–4297.

Morgan, M., S. Falcon, and R. Gentleman, 2018a *GSEABase: Gene set enrichment data structures and methods*. R package version 1.44.0.

Morgan, M., V. Obenchain, J. Hester, and H. Pagès, 2018b *SummarizedExperiment: SummarizedExperiment container*. R package version 1.12.0.

National Cancer Institute, 2019a PDQ Non-Small Cell Lung Cancer. https://www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq, Accessed: 2019-06-12.

National Cancer Institute, 2019b TCGA's Study of Lung Squamous Cell Carcinoma. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers/lung-squamous, Accessed: 2019-06-12.

Park, Y. R., S. H. Bae, W. Ji, E.-J. Seo, J. C. Lee, *et al.*, 2017 Gab2 amplification in squamous cell lung cancer of non-smokers. Journal of Korean medical science **32**: 1784–1791.

Pio, R. and L. M. Montuenga, 2009 Alternative splicing in lung cancer. Journal of Thoracic Oncology **4**: 674 – 678.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, *et al.*, 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**: e47.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140.

Stewart, B., C. P. Wild, *et al.*, 2014 World cancer report 2014. World Cancer Reports .

Weinstein, J. N., E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, *et al.*, 2013 The cancer genome atlas pan-cancer analysis project. Nature genetics **45**: 1113.

Xiao, Y., M. Feng, H. Ran, X. Han, and X. Li, 2018 Identification of key differentially expressed genes associated with non-small cell lung cancer by bioinformatics analyses. Molecular medicine reports **17**: 6379–6386.

Table 1 Highlighted enriched GO terms from differentially expressed genes

GOBPID[a]	P-value	Odds Ratio	Count[b]	Size[c]	Term
GO:0007172	1.14x10-3	27.78	4	5	Signal complex assembly
GO:0038063	6.44x10-3	9.26	4	7	Collagen-activated tyrosine kinase receptor signalling pathway
GO:0048025	7.26x10-3	4.63	6	15	Negative regulation of mRNA splicing, via spliceosome
GO:0006376	2.81x10-3	4.28	8	21	mRNA splice site selection
GO:0007189	5.21x10-4	2.64	20	73	Adenylate cyclase-activating G protein-coupled receptor signaling pathway

[a] Excerpt of the complete GO results table found under Supplementary Materials, with highlighted GO terms that were discussed.
[b] Number of genes mapped to the gene set
[c] Size of the gene set