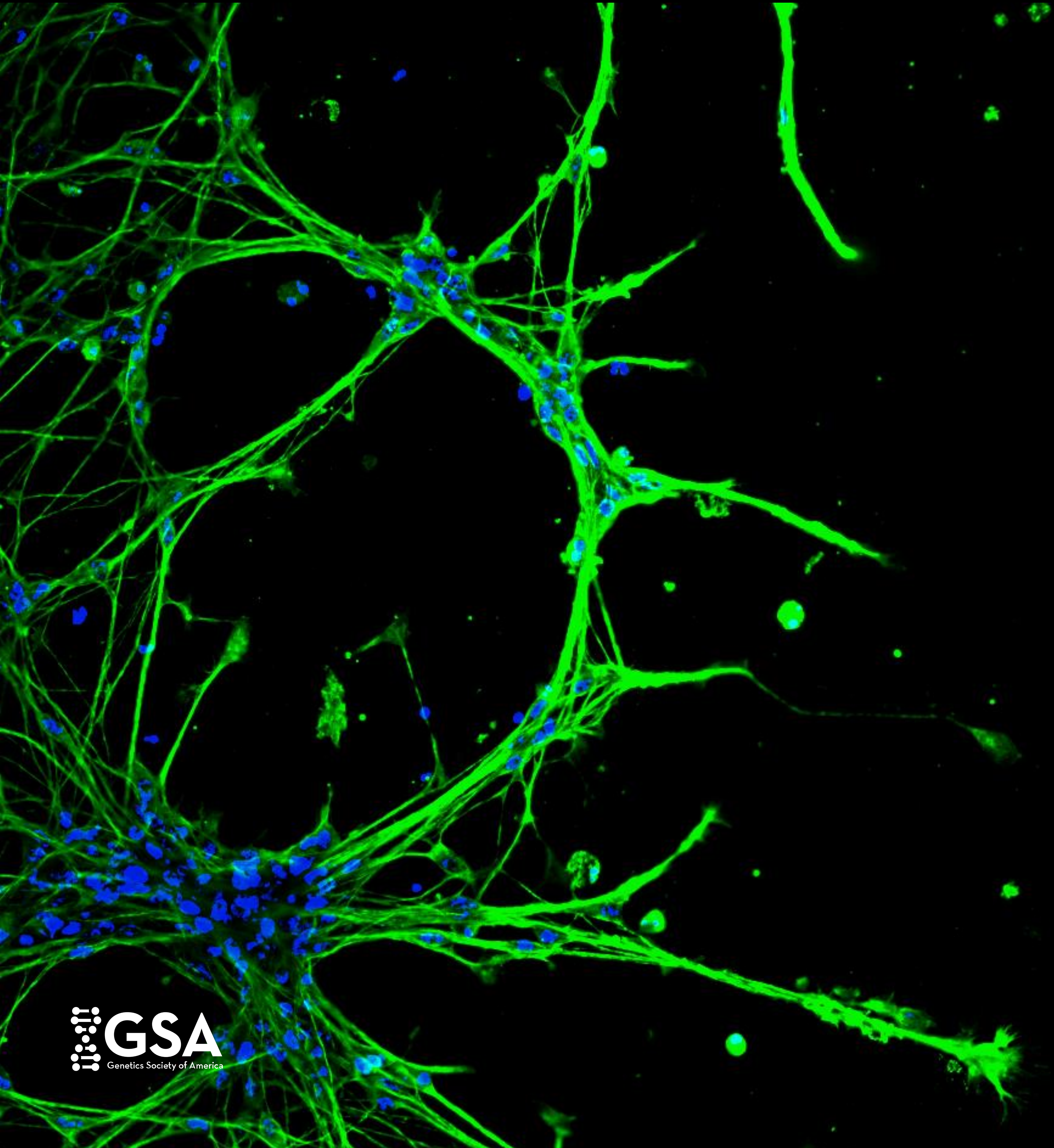
GENETICS

JANUARY 2022 · VOLUME 220 · GROUP 5



GSA
Genetics Society of America

OXFORD GENETICS

RNA-Seq Analysis in Bladder Cancer Patients

Emiliano Navarro-Garre[1, 2,†, *], Pol Bonet[1, 2,†], Alberto Carrasco[1, 2,†], Aina Comas[1, 2,†] and Sofía Redondo[1, 2,†]

[1]Departament de Genètica i de Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.
[2]Group 5, Current Topics in Bioinformatics Subject.
[†]These authors contributed equally to this work.

[*]Corresponding author: Emiliano Navarro-Garre, Departament de Genètica i de Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain. Mail: emiliano.navarro@autonoma.cat

Abstract

RNA-seq is a recent approach to carry out expression profiling using high-throughput sequencing technologies, being the preferred option to simultaneously measure the expression of tens of thousands of genes for multiple samples. In this study, we walk through a gene-level RNA-seq differential expression analysis using Bioconductor packages to find genes over- or under-expressed in bladder cancer patients, one of the types of cancer most affected by tobacco use, in an early and advanced tumoral stages, finding a certain number of genes that may be associated with the uncontrolled development of bladder tissue.

Keywords: RNA-seq; Bladder cancer; Statistics

Introduction

Bladder cancer is any of several types of cancer arising from the tissues of the urinary bladder, where the main symptoms are blood in the urine, painful urination and lower back pain and appear when epithelial cells that line the bladder become malignant. Most bladder cancers are diagnosed at an early stage, when the cancer is highly treatable. But even early-stage bladder cancers can come back after successful treatment, due to different risk factors such as smoking, family history, recurrent urinary tract infections, exposure to certain chemicals, or having certain mutation in the genes that are linked to bladder cancer (Board 2002).

In this study we have performed an **RNA-seq analysis**, in order to **detect deferentially expressed genes in bladder cancer patients in an early and advanced tumoral stage**, i.e. find genes over- or under-expressed in these cancer patients. RNA-seq is a recent approach to carry out expression profiling using high-throughput sequencing technologies, being the preferred option to simultaneously measure the expression of tens of thousands of genes for multiple samples (Coronado and Carrón 2021).

The aim of this study is to **compare the transcriptomic profile of patients in an early stage with ones in an advanced stage** of this cancer to find the differential expressed genes.

Methods

All the information about methods and steps has been extracted from the practice script (Coronado and Carrón 2021).

Packages and tools used

To carry out this study, different **Bioconductor** (Morgan 2021) packages have been used, as it has packages that support high-throughput sequencing data analysis, including RNA-seq. The packages that were used in this study include core packages maintained by the Bioconductor core team for importing and processing raw sequencing data and loading gene annotations:

- **SummarizedExperiment**: contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typicall represent genomic ranges of interest and the columns represent samples (Morgan *et al.* 2021).
- **DESeq2**: estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution (Love *et al.* 2014).
- **org.Hs.eg.db**: genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers (Carlson 2021).
- **biomaRt**: provides an interface to a growing collection of databases implementing the BioMart software suite, such as Ensembl (Durinck *et al.* 2009).
- **edgeR**: differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests (Robinson *et al.* 2010).
- **tweeDEseq**: differential expression analysis of RNA-seq using the Poisson-Tweedie family of distributions (Esnaola *et al.* 2013).
- **GOstats**: a variety of basic manipulation tools for graphs, hypothesis testing and other simple calculations (Falcon and Gentleman 2007).
- **tweeDEseqCountData**: RNA-seq count data employed to illustrate the use of the Poisson-Tweedie family of distributions with the tweeDEseq package (Gonzalez and Esnaola

2021).

- **annotate**: using R enviroments for annotation (Gentleman 2021).

In addition, **ggplot2** (Wickham 2016), **ggrepel** (Slowikowski 2021) and **MetBrewer** (Mills 2021) packages have been used in order to represent graphically the results.

Data description

The data is available (Data) in a **Ranged Summarized Experiment** (RSE) format, which is a matrix-like container where rows represent ranges of interest and columns represent samples and it is provided by Recount, which is an online resource consisting of RNA-seq gene and exon counts for different studies, including The Cancer Genome Atlas (TCGA).

For the case of our study, we have **58037 genes** initially in a cohort of 433 patients with bladder cancer, but we had to filter out 2 individuals due to missing information, so our working cohort was **431 patients**, being **138 early stage** bladder cancer individuals and **293 late stage** individuals.

Normalization

When we work with RNA-seq data, we need to normalize it, because of (1) the number of counts is related to sequencing depth and transcript length and (2) the number of counts is proportional to the mRNA expression level. So the normalization of the data try 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).

We have performed three different normalizations with three different methods (1) RPKM (Mortazavi *et al.* 2008), (2) TMM (Robinson and Oshlack 2010) and (3) the DESeq2 package's own normalization (Love *et al.* 2014), representing the first two normalizations next to the unnormalized data, but we have worked with the DESeq2 normalization.

More information about the normalization methods are present in the RMarkdown file.

Differential expression analysis

The R package DESeq2 allows us to test differential gene expression analysis based on the negative binomial distribution and the starting point is a count matrix with one row for each gene and one column for each sample, indicating the stage of the tumor stage (Love *et al.* 2014).

The plotMA function allows the graphical representation of the $\log_2$ fold-change over the mean of normalized counts for all the samples. In this way, we can find the genes that are significanlty overexpressed or underexpressed in late tumours, applying the following criteria: (1) adjusted p-value lower than 0.001 and (2) 10 $\log_2$ fold-change. This last criteria is very stringent.

Post RNA-seq analysis

After differential expression analysis, we performed a post RNA-seq analysis, where we visualized the data and performed an enrichment analysis.

Visualization: To view the data after RNA-seq analysis, we performed a **volcano plot** to identify changes in our data set, plotting significance vs. fold change on the y and x axes, respectively. In this case, the **negative $\log_{10}$ of the adjusted p-value** is on the y-axis, the **$\log_2$ fold-change** on the x-axis and each point is colored based on the tumor stage (early or late).

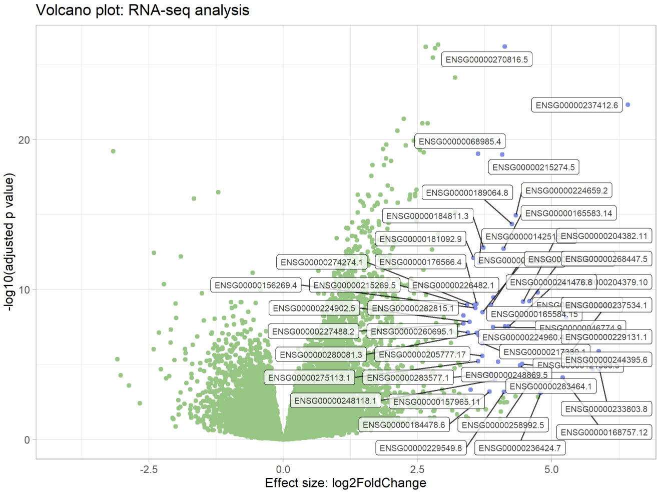


Figure 1 Volcano plot with the differently expressed genes in bladder cancer in blue color and the non-significantly differently expressed genes are colored in green.

The differential expressed genes are annotated with their ID and graphed using ggplot2 and ggrepel.

Enrichment analysis: The enrichment analysis was performed in order to **interpret gene expression**. This gene set enrichment analysis was based on the functional analysis of the differently expressed genes. This is useful for finding out if the differently expressed genes are associated with a certain biological process or molecular function. This analysis has been carried out using the biomaRt package.

Moreover, the enrichment results were represented with the ggplot2 package.

Results

After performing data normalization and differential expression analysis, we observed that **48 genes are over-expressed** and none under-expressed, depicted in Figure 1, where the genes differently expressed are colored in blue, while the non-significantly differently expressed genes are colored in green.

Once the differently expressed genes, in our case, over-expressed, have been obtained, the enrichment analysis has been carried out to find those processes most represented, as seen in Figure 2, where we can appreciate that the most represented Gene Ontology (GO) term is the **tissue development**, but without many differences with respect to the other GO terms in the chart.

Discussion

With this different RNA-seq expression analysis, we have found 48 differently expressed genes, all of them, over-expressed in late bladder cancer. These over-expressed genes may be oncogenes involved in the uncontrolled development of bladder tissue, but this hypothesis must be verified by searching for each of the over-expressed genes we have found in this report, due to the wide variety of associations with the GO terms.

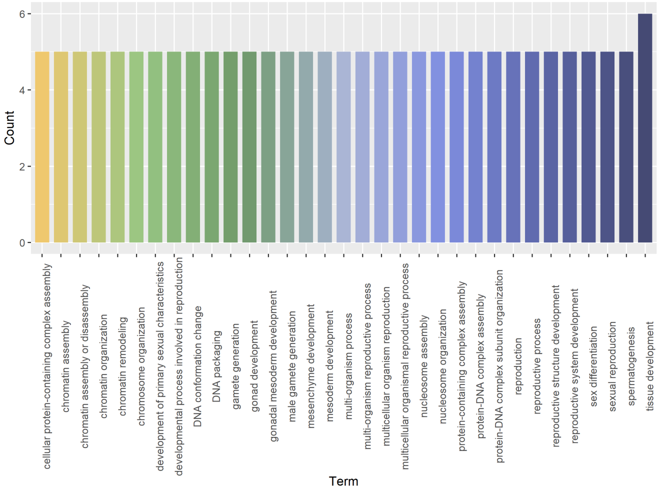


Figure 2 Gene ontology enrichment analysis of differently expressed genes in late bladder cancer.

Data availability

The data set used and the code to perform the analyses described in this article are presented below:

| Data.

| RMarkdown file.

| HTML file with the steps of the analysis.

Literature cited

Board PATE. 2002. Bladder cancer treatment (pdq®): Patient version. PDQ Cancer Information Summaries. .

Carlson M. 2021. *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.14.0.

Coronado M, Carrón N. 2021. Practice 4: Finding differentiantly expressed genes in cancer. Current Topics in Bioinformatics. .

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nature protocols. 4:1184–1191.

Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR. 2013. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated rna-seq experiments. BMC Bioinformatics. 14:254.

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

Gentleman R. 2021. *annotate: Annotation for microarrays*. R package version 1.72.0.

Gonzalez JR, Esnaola M. 2021. *tweeDEseqCountData: RNA-seq count data employed in the vignette of the tweeDEseq package*. R package version 1.32.0.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology. 15:550.

Mills BR. 2021. *MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art*. R package version 0.1.0.

Morgan M. 2021. *BiocManager: Access the Bioconductor Project Package Repository*. R package version 1.30.16.

Morgan M, Obenchain V, Hester J, Pagès H. 2021. *SummarizedExperiment: SummarizedExperiment container*. R package version 1.24.0.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by rna-seq. Nature Methods 2008 5:7. 5:621–628.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26:139–140.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biology. 11:1–9.

Slowikowski K. 2021. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*. R package version 0.9.1.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Appendix

Supplementary Figures

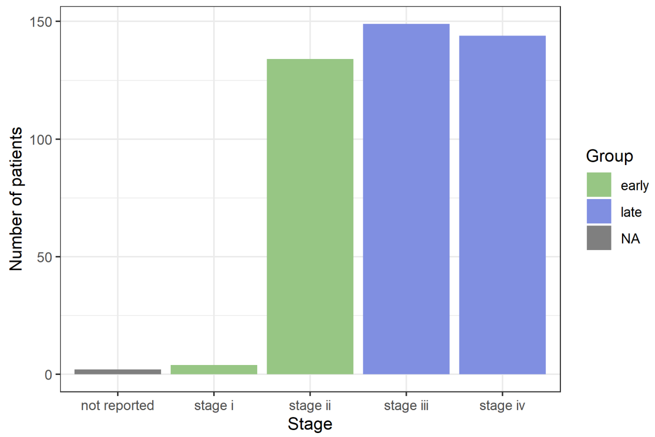


Figure 3 Number of patients in each tumoral stage.

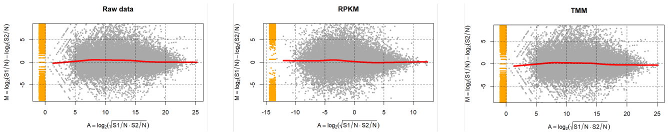


Figure 4 Representation of the 3 different methods to normalize counts.

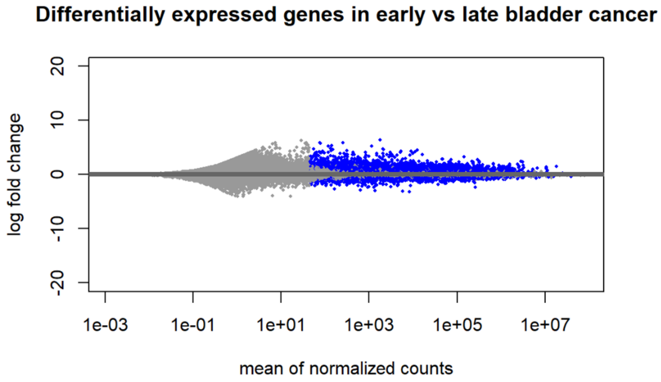


Figure 5 Differently expressed genes in early vs late bladder cancer

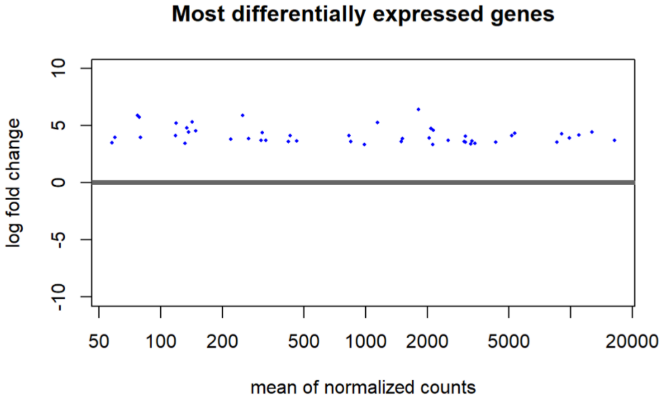


Figure 6 Most differently expressed genes.