Identification of differentialy expressed genes in colorrectal cancer using RNA-seq data

Miguel Román, Daniel Soto

February 2020

1 Abstract

Data visualisation, especially in genomics and molecular biology, is of particular interest. In this study, we analysed RNA-Seq data from colorrectal cancer (CRC) cells to find genes expressed differentialy, possibly involved in tumoral development and behavior.

2 Introduction

In this study, we analysed genes expressed differentialy in colorrectal cancer (CRC). For this purpose, we made use of the R software for statistical analysis and data visualisation.

3 Methods

The following R packages were used for this study[1][2]:

1. SummarizedExperiment: The SummarizedExperiment container contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the columns represent samples.

2. edgeR (Empirical Analysis of Digital Gene Expression Data in R): 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.

3. DESeq2 (Differential gene expression analysis based on the negative binomial distribution): 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.

4. tweeDEseq: RNA-seq data analysis using the Poisson-Tweedie family of distributions.
5. GOstats: A set of tools for interacting with GO and microarray data. A variety of basic manipulation tools for graphs, hypothesis testing and other simple calculations.
6. annotate: Using R enviroments for annotation for microarrays.
7. org.Hs.eg.db: Genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers.
8. biomaRt: Interface to BioMart databases (i.e. Ensembl).
9. ggplot2: ggplot2 is a system for declaratively creating graphics.
10. ggrepel: Provides text and label geoms for 'ggplot2' that help to avoid overlapping text labels. Labels repel away from each other and away from the data points.

First of all, we plotted a histogram that shows the count of CRC tumor classified by the stage of development. Results can be seen in figure 1

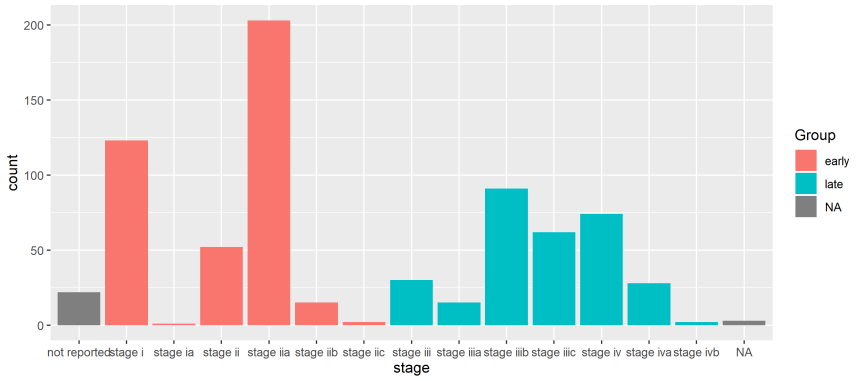

Figure 1: CRC tumor stage count

After that, as depicted in section "results", we represented in a volcano plot the analysed genes to explore which of them are candidates of being involved in cancer development. A volcano plot is a type of scatter plot (points) that is used to identify changes in large data sets composed of replicate data. It plots significance versus fold-change on the y and x axes, respectively.[3]

4 Results

In figure 2 it is shown that some genes are differentialy expressed in CRC cells. Nonetheless, since these genes do not appear in the corresponding gene ontology, results are hard to explain.

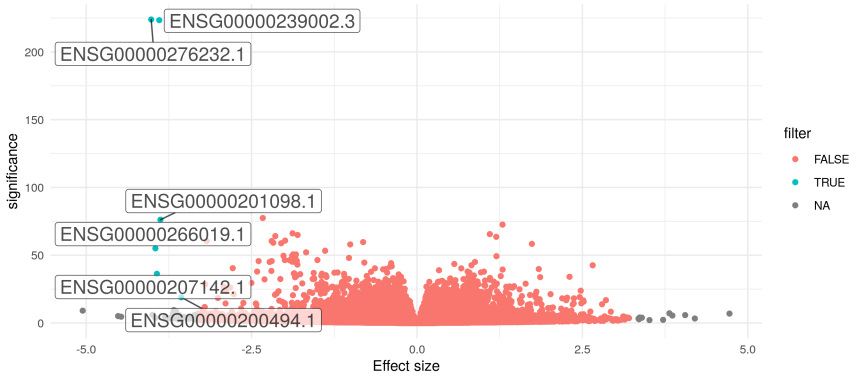

Figure 2: Volcano plot showing statistically significant differentialy expressed genes in CRC cells.

References

[1] Bioconductor. https://bioconductor.org/). Accessed: 2020-02-06.

[2] R documentation. https://www.r-project.org/other-docs.html). Accessed: 2020-02-06.

[3] Wikipedia. https://en.wikipedia.org/wiki/Volcano$_p lot_($ $statistics$).$Accessed$: $2020 - 02 - 06$.