Temes Actuals de la Bioinformàtica

Pages 1–4
2022

Finding differentially expressed genes in stomach cancer

Gisela de Miguel Garcia, Olivia Dove Estrella, Oriol Gómez Lores, Roger Parramon Codina

Universitat Autònoma de Barcelona

January 2022

Keywords: stomach cancer, RNA-seq, expression analysis, gene ontology

ABSTRACT

Stomach cancer, also called gastric cancer, is the fifth most common and the third most lethal neoplasm in the world [1]. Stomach cancer is more common in East Asia, Eastern Europe, and South America [2]. There are several factors that increase the risk of stomach cancer, such as obesity, a diet high in salty and smoked foods or low in fruits and vegetables, family history of stomach cancer, infection with *Helicobacter pylori*, long-term stomach inflammation, among others [2].
Most stomach cancers are adenocarcinomas (95%) [3]. Other types of gastric cancer include lymphomas, gastric sarcomas and neuroendocrine tumours, but these are less common. There are five stages of stomach adenocarcinoma. Early tumours are those in stages *i* and *ii*, while late tumours are those in stages *iii* and *iv*. The stage of a cancer informs how big the tumour is and whether it has metastasised [4]. It is important to determine the stage in which stage the tumor is, this helps doctors to decide the best course of treatment.
Here we conducted an RNA-seq expression computational analysis on RNA-seq data from 427 patients diagnosed with stomach cancer to compare the transcriptomic profile of patients in early tumoral stage with the ones in an advanced stage of this disease. We found 17 differentially expressed genes that were under-expressed in late-stage tumours.

CONTENTS

METHODS

In this section, we provide the packages and tools used, a data description as well as the statistical tests performed.

Packages and tools used

All the analyses were conducted using **R version 4.1.2** [5]. Several packages from this programming language were used. To perform the gene ontology enrichment study, we employed the **Ensembl** database [6].

- **ggplot2**: R data visualization package.
- **ggrepel**: R package that builds on ggplot2, allowing better data labelling in graphs.
- **BiocManager**: R package that enables Bioconductor software, which helps with the analysis of biological data.
- **DESeq2**: R package from Bioconductor which tests differential gene expression analysis based on the negative binomial distribution.
- **biomaRt**: R package from Bioconductor that finds information from Ensembl database on the differentially expressed genes.
- **GOstats**: R package from Bioconductor that tests the association of Gene Ontology (GO) terms to genes in a gene list.

Data description

Our data comes from Recount, an online resource consisting of RNA-seq gene and exon counts for different studies, including the Cancer Genome Atlas data. The original dataset contained 58037 genes analyzed in a cohort of 453 patients with stomach cancer. Specifically, 206 patients are classified in early tumoral stages, whereas 221 patients are included in late tumoral stages as seen in Figure 1. Also, there were 26 patients whose information regarding their tumoral stage was not provided. These patients were discharged from our analysis, thus, 427 patients were finally included for the investigation.

Apart from the tumoral stage, our dataset also contained information regarding the read counts data and the phenotype of the patients. We also checked that the same individuals were found in both the counts dataset and in the phenotype dataset. We found that the same individuals were in both datasets, therefore, no individual was discharged.
In Table 1 a summary of the information from our data is provided.

Table 1. Individuals and gene data summary.

Total number of individuals original dataset	453
Total number of individuals after filtering	427 (26 patients removed)
Total number of individuals in early stages	206
Total number of individuals in late stages	221
Total number of genes	58037

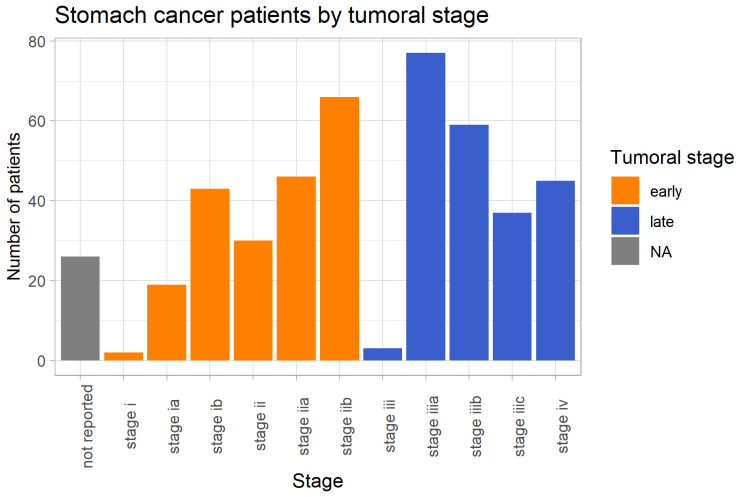

Figure 1. Stomach cancer patients by tumoral stage

Statistical tests

To perform the differential expression analysis we used the DESeq2 R package, which is very convenient because it includes its own normalization method. With this analysis, it is possible to find the genes that are significantly over-expressed or under-expressed in late tumours (Figure 2). In order to keep the most significant and biologically relevant genes, those genes which had an adjusted p-value lower than 0.001 and had a $10 log2 fold - change$ were selected. After applying these criteria, only 17 were kept (Figure 3). The commented code is available in our GitHub repository.

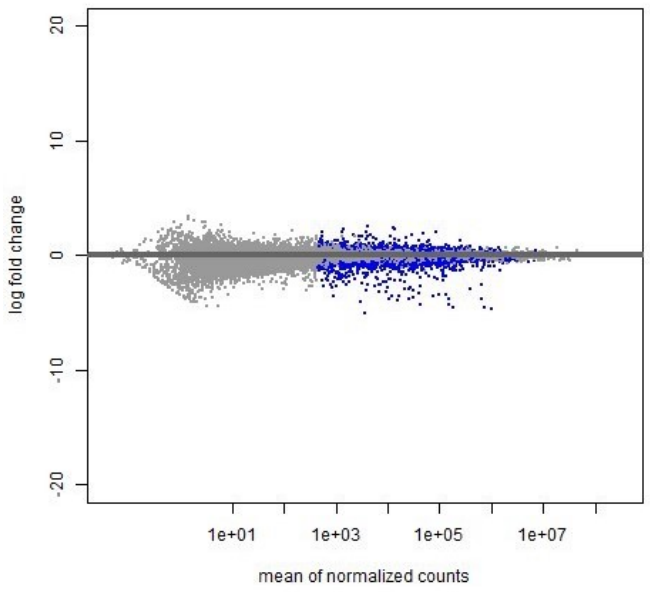

Figure 2. Differentially expressed genes in early vs late stomach cancer. Points coloured in blue have an adjusted p-value lower than 0.1.

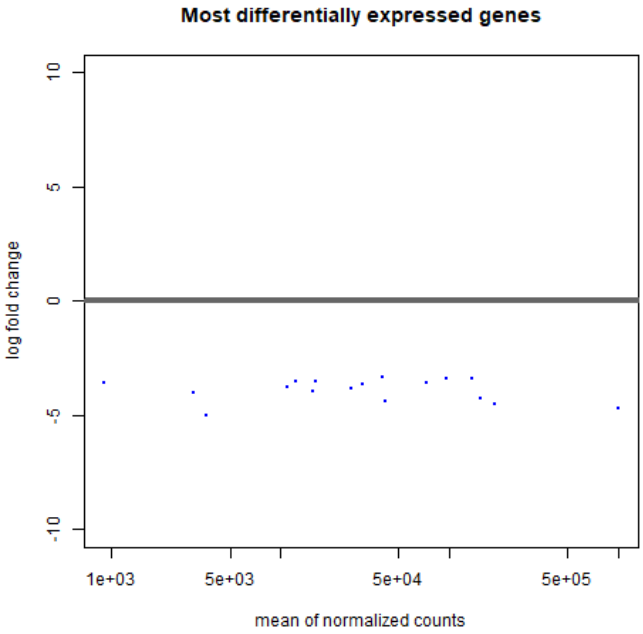

Figure 3. Representation of the most differentially expressed genes.

RESULTS

Here we present the results from the RNA-seq analysis and from the Gene Ontology enrichment analysis.

Visualization of RNA-seq analysis

When we represent only the most diferentially expressed genes (Figure 3), all 17 retrieved genes were under-expressed (see Table 2 for more information regarding these genes).

Table 2. Description of the most differentially expressed genes.

Ensembl ID	Entrez ID	Symbol	Chromosome
ENSG00000016602	22802	*CLCA4*	1
ENSG00000125780	7053	*TGM3*	20
ENSG00000136694	27179	*IL36A*	2
ENSG00000140519	51458	*RHCG*	15
ENSG00000163207	3713	*IVL*	1
ENSG00000169509	54544	*CRCT1*	1
ENSG00000170423	196374	*KRT78*	12
ENSG00000171401	3860	*KRT13*	17
ENSG00000185873	132724	*TMPRSS11B*	4
ENSG00000187054	339967	*TMPRSS11A*	4
ENSG00000196805	6701	*SPRR2B*	1
ENSG00000203785	6704	*SPRR2E*	1
ENSG00000203786	448834	*KPRP*	1
ENSG00000204544	394263	*MUC21*	6
ENSG00000229035	-	*SPRR2D*	1
ENSG00000229732	-	-	17
ENSG00000241794	6700	*SPRR2A*	1

The following volcano plot (Figure 4) aims to represent the results derived from the RNA-seq analysis.

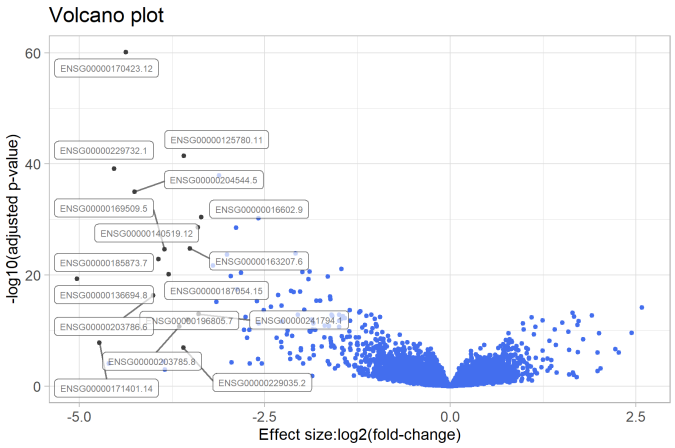

Figure 4. Volcano plot representing the results from the RNA-seq analysis. Points coloured black correspond to those 17 genes which are most deferentially expressed.

Enrichment analysis

As the final step to interpret gene expression data, we performed a gene set enrichment analysis based on the functional annotation of the differentially expressed genes (Figure 5). This is useful for finding out if the differentially expressed genes are associated with a certain biological process or molecular function.

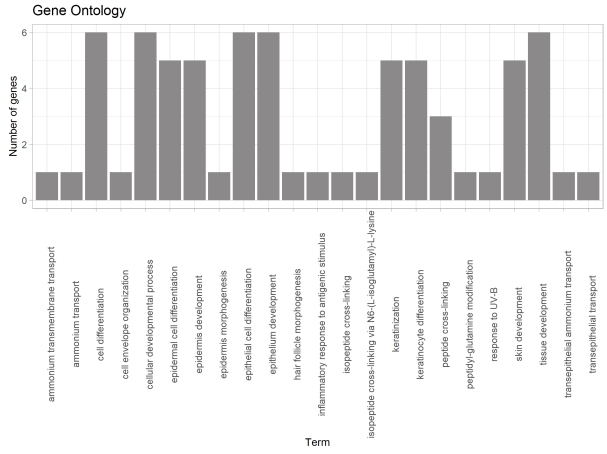

Figure 5. Functional annotation of the most differentially expressed genes.

The most prevalent gene functions among the differentially expressed genes in patients with stomach cancer are mainly related to development and cellular differentiation, especially tissue differentiation, with epithelial tissue being the most relevant.

DISCUSSION

Interestingly all 17 differentially expressed genes are underexpressed in late-stage tumours compared to the early-stage ones. We can link this fact to the gene functions determined by the Gene Ontology enrichment analysis, where cellular differentiation was the most common function. Three of the most important processes in a cell are proliferation, apoptosis, and differentiation, and all three of them are very relevant in cancer evolution [7]. While cellular differentiation is not completely understood yet, the relevance it has in neoplastic evolution is undeniable. The lack of cellular differentiation, called anaplasia, is usually considered a typical sign of cancer. The dedifferentiation process of specialized cells, where they lose their function and structure (and acquire certain stem cell-like properties) [8], can result in an increase of unregulated growth that would be impossible for healthy cells, with growth regulation pathways active [9]. In this specific case, it could be that the low levels of expression of such genes, in a tissue very exposed to friction, mechanical wearing and thus in need of constant cell renovation, significantly increases the likelihood of a tumour appearing. Even if cellular dedifferentiation is a hallmark of tumour growth in most cases, in the stomach the constant need for new tissue could result in a massive proliferation of the altered cells. The opposite could be true, meaning that the comparatively high rate at which cells are replaced in the stomach results in an increased likelihood of mutation due to high numbers of cell divisions.

REFERENCES

[1] Cuzzuol BR, Vieira ES, Araújo GRL, Apolonio JS, de Carvalho LS, da Silva Junior RT, et al. Gastric cancer: A brief review, from risk factors to treatment. *Archives of Gastroenterology Research*, 1(2):34–39, 2020.

[2] Mayo Clinic. Stomach cancer. Retreived from: https://www.mayoclinic.org/diseases-conditions/stomach-cancer/symptoms-causes/syc-20352438, 2021. Accessed: 2021-12-30.

[3] American Cancer Society. Stomach cancer risk factors. Retreived from: https://www.cancer.org/cancer/stomach-cancer/causes-risks-prevention/risk-factors.html, 2021. Accessed: 2021-12-30.

[4] Cancer research UK. Stomach cancer. Retreived from: https://www.cancerresearchuk.org/about-cancer/stomach-cancer, 2019. Accessed: 2021-12-30.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[6] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode RM, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, et al. Ensembl 2021. *Nucleic Acids Research*, 49(1):884–891, 2020.

[7] Enane FO, Saunthararajah Y, and Korc M. Differentiation therapy and the mechanisms that terminate cancer cell proliferation without harming normal cells. *Cell Death Disease*, 9(9):1–15, 2019.

[8] Kin Fong L, Yu-Chen H, Chia-Hao H, Chun-Hao H, and Ping Ching P. Characterization of stem cell-like property in cancer cells based on single-cell impedance measurement in a microfluidic platform. *Archives of Gastroenterology Research*, 1(2):34–39, 2020.

[9] Norris A and Korc M. Chapter 324 - aberrant signaling pathways in pancreatic cancer: Opportunities for targeted therapeutics. In Bradshaw RA and Dennis EA, editors, *Handbook of Cell Signaling (Second Edition)*, pages 2783–2798. Academic Press, San Diego, second edition edition, 2010.