BIOINFORMATICS ANALYSIS IN GENE EXPRESSION DATA FROM PUBLICLY AVAILABLE DATABASES

SARTOR, IVAINE TAÍS SAUTHIER[1,2]; RECAMONDE-MENDOZA, MARIANA[3]; ASHTON-PROLLA, PATRICIA[1-4]
[1] Laboratório de Medicina Genômica, Serviço de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, LMG-SPE-HCPA; [2] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, PPGBM-UFRGS;[3] Departamento de Informática Aplicada, Instituto de Informática, Universidade Federal do Rio Grande do Sul, INF-UFRGS; [4] Departamento de Genética, Universidade Federal do Rio Grande do Sul, DEGEN-UFRGS.

The gastrointestinal (GI) cancers account for 20% of estimated new cancer cases and 15% of estimated death worldwide. Among GI neoplasias, we especially highlight: i) the esophageal cancer (EC), which comprises two subtypes: the squamous cell carcinoma (SCC) and adenocarcinoma (ADC), the last one can progress from a pre-malignant lesion, known as Barrett's esophagus; ii) the gastric cancer (GC), which can progress from pre-malignant lesions such as chronic gastritis (ChG) and intestinal metaplasia (IM) and iii) the colorectal cancer (CRC), which can be stablished from colorectal adenoma lesion. GI malignancies are aggressive and heterogeneous diseases with poor survival. Further knowledge about the molecular pathogenesis and biological features of GI cancers is necessary to enable the identification and characterization of novel molecular biomarkers and therapeutic targets. In a previous study, which used a computational approach, was identified the transcription factor *TULP3* as a master regulator of carcinogenesis in pancreatic ductal adenocarcinoma (PDAC). The authors observed a poor prognosis in patients with higher *TULP3* expression in PDAC. Considering that pancreas and other gastrointestinal organs (such as esophagus, stomach and intestine) have the same embryonic origin, we investigated the profile of *TULP3* gene expression in GI tissues hypothesizing that it may have a role in these diseases. Therefore, we performed bioinformatics analysis to compare *TULP3* expression in GI tissues and to analyze patient survival. Gene expression data from patient biopsies were obtained from GEOdatabase and TCGA public repositories. GEOdatasets were downloaded under accession numbers: GSE26886 (GPL570) and GSE1420 (GPL96) for EC; GSE79973 (GPL570), GSE33335 (GPL5175) and GSE2669 (GPL2048) for GC; and GSE21510 (GPL570) and GSE24514 (GPL96) for CRC. From TCGA we obtained the RNASeq data of the following studies: ESCA (Esophageal Carcinoma), STAD (Stomach Adenocarcinoma) and READ (Rectum Adenocarcinoma). Preprocessed microarray data from COAD-TCGA (Colon Adenocarcinoma) study was obtained as provided by the authors. Principal component analysis was performed in each study to filter possibly biased samples. Gene expression raw data were normalized using *affy BioConductor* R-package for GPL570 microarrays and *oligo BioConductor* R-package for GPL5175 and GPL2048 microarrays. Raw counts from RNASeq data were normalized using *limma BioConductor* R-package. To select a single probe to represent a gene, we used the JetSet score for Affymetrix GPL570 and GPL96 microarrays. Data normality assumptions were verified and the appropriate statistical tests were chosen. Survival analysis was performed using *survival* R-package, the Kaplan-Meier method was used to estimate survival curves and LogRank test was used to compare the curves. TULP3 gene expression comparison between groups in ESCA-TCGA (*p*-value=3.21e-06), GSE26886 (*p*-value=2.03e-06) and GSE1420 (*p*-value=0.01) could differentiate the esophageal lesions. Despite *TULP3* showed significant statistical differences, the esophageal lesions analyzed and the expression trend observed in all datasets were not the same. Nevertheless, the survival analysis in TCGA Esophageal ADC associated a poor prognosis in patients with higher *TULP3* expression, ranging from ($\log_2$4.45, $\log_2$6.16], with a *p*-value=0.03, HR=2.11(1.05-4.21), while in TCGA Esophageal SCC, an unfavorable prognosis was associated with lower *TULP3* expression ($\log_2$3.62, $\log_2$5.34], *p*-value=0.04, HR=0.46(0.22-0.94). Considering GC, *TULP3* analysis in STAD-TCGA (*p*-value=0.02), GSE33335 (*p*-value=4.45e-07) and GSE2669 (*p*-value=3.44e-05) presented higher expression in gastric cancer samples in comparison with adjacent non-tumoral mucosa (non-GC) and ChG and IM. When we analyzed survival probability according the gender of patients in STAD-TCGA study we observed a worse prognosis in females with higher *TULP3* levels, ranging from ($\log_2$5.07, $\log_2$7], with a *p*-value=3.77e-3, HR=2.44(1.30-4.44). In male patients no difference was observed. In addition, increased *TULP3* expression in diffuse-type GC was also associated to an unfavorable prognosis ($\log_2$4.91, $\log_2$6.25], *p*-value=0.04, HR=2.93(1.00-8.54), but the same trend was not observed in the group of patients with intestinal-type GC. In patients diagnosed with gastric cancer not otherwise specified (NOS), higher *TULP3* expression ($\log_2$5.05, $\log_2$7], was also associated with worse prognosis, *p*-value=2.29e-04, HR=3.46 (1.71-6.98). The dichotomized TULP3 expression presented significant difference in univariate analysis in diffuse-type (*p*-value=0.03; HR=2.93(1.00-8.54)) and NOS (*p*-value=3.34e-04; HR=3.46(1.71-6.98)). Although the *TULP3* gene expression analysis in EC samples showed significant differences, the lesions analyzed and the observed trend of its expression in all datasets were not the same. In addition, the prognostic value associated to esophageal ADC and SCC, despite statistical significance, should be exploited in future works to comprehend biological process involved in EC. Considering GC, higher *TULP3* gene expression was observed in GC groups in STAD and GSE33335 studies, and a worse prognosis was associated with higher *TULP3* expression. Finally, in CRC higher *TULP3* gene expression was observed in CRC group in all studies and poor prognosis was assigned in patients with lower expression. Indeed, it is possible that *TULP3* has a role as a biomarker, and more studies are needed to confirm these *in silico* findings.

Keywords: gastrointestinal cancers, gene expression, bioinformatics