Oncology 2025
The Cancer Control Challenge:
Better Understanding It to Better Face It

DETECTION OF COLORECTAL CANCER BIOMARKERS FROM TUMOR-EDUCATED PLATELET RNA USING MACHINE LEARNING

Felipe Mateus Souza Serrão[1], Juscelino Carvalho de Azevedo Junior[2], Guilherme Cardoso Almada[3], Ana Beatriz Lima Belicha[2], Valéria Cristiane Santos da Silva[2], Danielle Queiroz Calcagno[2].

[1] Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil.
[2] Núcleo de Pesquisas em Oncologia, Universidade Federal do Pará, Belém, Brazil.
[3] Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil.

Introduction: Colorectal cancer (CRC) is one of the most common gastrointestinal malignancies worldwide and represents a relevant public health problem. Patients are usually diagnosed at advanced stages due to the absence of symptoms in the early stages of the disease. Thus, tumor-educated platelets (TEPs) have emerged as a minimally invasive liquid biopsy tool for the detection of various types of tumors, including CRC. **Objectives:** Therefore, we aimed to identify a molecular signature in TEPs for the diagnosis of CRC, as well as to determine the differentially expressed genes and compare two different machine learning decision tree-based algorithms for CRC detection. **Methods**: We downloaded platelet gene expression data from the Gene Expression Omnibus (GEO; GSE183635), containing 326 samples from CRC patients and 67 samples from control subjects. Low-quality reads were filtered using FastP, and the remaining high-quality reads were aligned to the human genome (GRCh38.p13 v43) using the Salmon tool. Aligned reads were then imported into RStudio via the Tximport package for further analysis. We performed differential expression analysis using DESeq2 ($|log2FoldChange| > 1$, adjusted p-value < 0.05). Gene ontology (GO) enrichment analysis (p-value < 0.05) was conducted on differentially expressed genes (DEGs). DEGs were used as input for two algorithms: randomForest and XGBoost. The model with the highest AUC was selected as the best performer, allowing the identification of relevant genes. **Results**: Differential analysis showed that 7,177 DEGs were found, in which 55 (0.77%) were upregulated and 7,122 (99.23%) were downregulated in CRC patients compared to controls. Upregulated genes are involved in oxygen transport and oxidative stress processes, while downregulated genes are related to immune system regulation and cell differentiation. XGBoost outperformed randomForest (AUC 0.848 vs. 0.788). The upregulated genes considered relevant were *MYL9* (AUC = 0.806 and CI = 0.748-0.864), *IGFBP2* (AUC = 0.661 and CI = 0.585-0.737) and *ALAS2* (AUC = 0.674 and CI = 0.599-0.749), with *MYL9* standing out with the highest AUC value. **Conclusion:** Our results reveal the potential of *MYL9* expression in TEP cells as a diagnostic biomarker

for CRC, reinforcing the potential of TEP cells as a minimally invasive source of biomarkers.

Keywords: precision medicine; liquid biopsy; tumor-educated platelets; machine learning; gene expression.