Gene embedding: A novel machine learning approach to identify gene candidates related to immunotherapy responsiveness

C.T. Choy1, C.H. Wong2, S.L. Chan2
1Department of Clinical Oncology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China, 2The Cancer Drug Testing Unit, State Key Laboratory of Oncology in South China, Department of Clinical Oncology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China

Background: Apart from PD-L1 and mutational load, there are no genetic predictive biomarkers for checkpoint inhibitors treatment. In this study, gene embedding, a machine learning technique, was used to single out related genes of immune checkpoint proteins (i.e. PD-1, PD-L1, CTLA-4) as new potential predictors for such responders.

Methods: TCGA RNASeqV2 level 3 RSEM normalized read counts (January 2016) were downloaded from the Broad Institute TCGA GDAC Firehose. A shallow neural network, aka embedding layers for samples and genes, were trained using log2 transformed data. Neighbors closeness were evaluated by euclidean distance. The model was kept blind from any additional information, including cancer types, protein-protein interactions and gene ontologies.

Results: Gene expressions of 13045 samples from 36 cancer types were embedded into 50-dimension space, while cancer types were learnt by the model without supervision. Immunotherapy responders and non-responders were stimulated from melanoma (SKMC) and lung squamous cell carcinoma (LUSC) data, and hepatocellular carcinoma and prostate cancer data respectively. 9 genes (TNFRSF8, CLREC10A, FCN1, CD8B, SLA2, IL2RA, CTLA4, GZMH), 3 genes, and 6 (SH2D1A, MEI1, PDCD1, GIF1, ST1, SIRPG) genes were found to be closely related neighbors with PD-1, PD-L1, and CTLA-4 respectively in responders but not in non-responders. All neighbors were neither co-expressed in SKMC/LUSC dataset nor indicated as interacting partners on existing databases (BioGRID, MINT, iRefWeb, STRING, HPRD and Reactome). 88.8% genes were evidenced as either directly related or connected to checkpoint proteins and/or T cells activation in literature. Further evaluation of the role of identified targets in immune checkpoint blockade therapy would be warranted.

Conclusions: We identified potential biomarker candidates for immune checkpoint blockade therapy by TCGA data mining and demonstrated the utility of gene embedding learned from big gene expression dataset as a powerful tool to uncover gene relationships that may not be discovered otherwise without prior knowledge on functional interactions.

Legal entity responsible for the study: Department of Clinical Oncology, Faculty of Medicine, The Chinese University of Hong Kong.

Funding: Department of Clinical Oncology, Faculty of Medicine, The Chinese University of Hong Kong.

Disclosure: All authors have declared no conflicts of interest.