Leveraging public pan-cancer transcriptomic data for immuno-oncology: systematic collection and a graphical tool for analysis

Marco Barreca[1,2], Eduardo Aranda Cañada[3], Immuno-model Data Gathering Taskforce[4], Dr. Rebeca Sanz Pamplona[3,5], Maurizio Callari[2]

[1] *Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy*
[2] *Fondazione Michelangelo, Milan, Italy*
[3] *Cancer Heterogeneity and Immunomics (CHI), University Hospital Lozano Blesa, Aragon Health Research Institute (IISA), Zaragoza, Spain.*
[4] *Members of Immuno-model COST Action CA21135*
[5] *ARAID Foundation, Aragon Government, Zaragoza, Spain*

Development of omics technologies has generated a growing amount of publicly available molecular profiles. Such goldmine could be interrogated to generate or validate hypotheses on new biomarkers or therapeutic targets. However, access and analysis of such datasets, in particular for non-computational experts, could be challenging. Hence, the goals of this work were to: i) implement a Systematic Search Strategy to collect publicly available transcriptomics datasets from immuno-oncology related experiments, ii) facilitate access to such resource by developing a graphical interface for the analysis.

The project was divided into a pilot and main phase. The first focused on setting the basis of the Systematic Search Strategy, i) collecting information to define the search boundaries, ii) developing a code to interrogate the data repositories and collect the query results and iii) determining the criteria for manual curation and dataset selection by expert researchers. In the main phase, queries were run on the two selected repositories (Gene Expression Omnibus and ArrayExpress) and the query results are being evaluated by the members of the established taskforce to identify datasets matching the selection criteria. To allow the graphical-based data mining and multiple unsupervised and supervised analyses, a Shiny app is under development.

The pilot phase highlighted that a manageable number of candidate datasets per cancer type could be found, without needing to narrow down the search with specific keywords. Evaluation by multiple researchers of the same query results emphasized the need for clear selection criteria to reduce discrepancies. Such criteria could be summarized in: bulk or single-cell transcriptomic datasets from pre-clinical or clinical samples receiving immunotherapy, alone or in combination with other treatments.

In the main phase, queries were run for over 20 distinct cancer types. Candidate datasets collected with each query are being evaluated by at least two expert researchers. Statistics to thoroughly describe the collected datasets will be computed and presented at the meeting. The collected datasets will be made available for analysis through an online open-access and user-friendly graphical tool. Examples of the tool capabilities will be presented.

This project is contributing to create knowledge-sharing channels that could foster the connection between basic, translational, and clinical investigators, supporting the development, optimization and use of preclinical models for immunotherapy research.

This work was supported by Immuno-model COST Action VMG, in the extended call from September 23rd to October 21st, 2024.