**TACITuS: Transcriptomic Data Collector, Integrator, and Selector on Big Data Platform**

**Alaimo S(1), Di Maria A(2), Shasha D(3), Ferro A(1), and Pulvirenti A(1)**

1. **Department of Clinical and Experimental Medicine, University of Catania, Catania, IT**
2. **Department od Physics and Astronomy, University of Catania, Catania, IT**

**Courant Institute of Mathematical Science, New York University, NY, USA**

**MOTIVATION**

Transcriptome analysis helps to understand the dynamics regulating biological and pathological processes. Gene expression can be used to define accurate biomarkers in precision medicine, or to describe how observed alterations impact on the patient's phenotype through the evaluation of biological pathways. To date, many studies are moving in this direction, producing a huge amount of data stored in databases such as NCBI GEO or ArrayExpress. However, integrating such information, while desireable for large-scale studies, is not always easy, due to non-uniform formats and platform heterogeneity, among other issues..

**METHODS**

Here we present TACITuS (Transcriptomic Data Collector, Integrator, and Selector) a web portal that simplifies the process of collection, pre-processing, selection, and integration of transcriptomics data. Through our interface a user can collect data from major sources, such as NCBI GEO or ArrayExpress, and integrate them with their own data in a standardized format, facilitating subsequent analyzes. Our tool, thanks to Apache SPARK, MongoDB, Apache Lucene, and proper indexing algorithm, can easily manage large amounts of data guaranteeing suitable performances.

Furthermore, all metadata are collected and standardized enabling fast search and easy management on large datasets. Through the integration of user data with several heterogeneous sources, our database will facilitate subsequent analyses, generating more robust results, due to the increase of the sample base from which hypotheses are generated. Our software is freely available at https://tacitus.alaimos.com/ and distributed through an Open-Source license. Finally, connection with computational platforms such as Galaxy will be provided to enable complex data analysis.

**RESULTS**

Simplified filtering and integration of transcriptomics data is a crucial step for biomarker detection and precision medicine pipelines. To evaluate our methods, we imported several high dimensional transcriptomics datasets to determine efficiency in storage, pre-processing, and indexing. On such datasets, we also performed several selection queries and gathered their results to demonstrate the user experience. In a few minutes the user can select and process big datasets and then prepare the results for subsequent analysis (Figure 1).

