Integration of GMQL within the Galaxy and Taverna frameworks

**Background:** Novel technologies for high-throughput DNA sequencing, namely Next Generation Sequencing (NGS) methods [1], are deeply changing the way in which biological experiments are conducted, by producing data at increasing speed and reducing costs; therefore, managing NGS data is quickly becoming the biggest "big data" problem of mankind.

In this scenario, the Bioinformatics group at Politecnico di Milano has been developing the *GenoMetric Query Language* (GMQL) [2]. GMQL is a tool that provides a next-generation query language for querying NGS data, thus supporting biologists in analyzing NGS data to discover novel genomic properties. The added value of GMQL is its scalability and ability to run computations in the cloud; since it is implemented over Hadoop [3], it is suitable for processing Tera Bytes of genomic data, such as the ones provided by ENCODE [4] and TCGA [5] consortia.

Currently, GMQL is used in several joint projects, which involve studies on cancer and genetic diseases and are carried on with the European Institute of Oncology (IEO) [6] and the Center for Genomic Science of the Istituto Italiano di Tecnologia (IIT) [7] at the IFOM-IEO campus [8]. GMQL is also part of the GenData 2020 [9] project, which involves nine Italian Universities. See currently available web interfaces for GMQL use at [http://cru.genomics.iit.it/GMQL/queries/](http://cru.genomics.iit.it/GMQL/queries/) and [http://131.175.120.18/gmql-services-webapp-2.0.2/](http://131.175.120.18/gmql-services-webapp-2.0.2/)

**Objectives:** This project requires integrating GMQL within the two most popular workflow manager systems for Life Sciences, i.e. *Galaxy* [10] and *Taverna* [11]. The aim is to make possible to include GMQL processing into complex analysis workflows, by creating a GMQL workflow element that takes any GMQL query as an input parameter. In the case of Galaxy, this has to make possible to deposit GMQL into the Galaxy ToolShed [12], i.e. the repository of Galaxy tools, and run it on a public instance of a public Galaxy server [13]. Tutorials are available about how to publish a tool on a Galaxy instance [13, 14]; using Docker containers [15] could be an option. Testing of the enabled use of GMQL in Galaxy and Taverna within real research workflows has also to be performed.

**Contacts:**

marco.masseroli@polimi.it, pietro.pinoli@polimi.it

**References:**

[6] European Institute of Oncology (IEO) [https://www.ieo.it/en/](https://www.ieo.it/en/)
[12] Tool shed [https://toolshed.g2.bx.psu.edu/](https://toolshed.g2.bx.psu.edu/)