# Project Proposal

**Title**

A Cloud computing approach for cancer genomics

**Project Description**

BACKGROUND

In the last decades, the rapid evolution of high-throughput NGS (Next Generation Sequencing) techniques reformulated the scope of genomics research, in the context of both physiological and pathological conditions. Moreover, the substantial decrease in the overall costs of NGS-based technologies led to previously unintended applications in several scientific fields, including disease characterization and drug discovery[1]. However, the exponential growth of NGS-derived biological data has posed the problem of how to address computational issues that arise from storing, transferring and analyzing huge amounts of sequenced data. In the next few years, petabytes of genomic data will be produced and, most importantly, stored in a system that nowadays is characterized by a centralized repository from which any researchers can download copies of the data. This flux of data, the so-called "bringing the data to the people" approach, cannot possibly scale in view of the massive growth in datasets that we are expecting in Life Sciences data.

MAIN GOAL

In this view, this project leverages a Cloud-based storage system, which flips the traditional approach in data sharing and processing. The solution is to change the paradigm and "bring the people to the data"[2], and, if applied to the majority of produced data, common standards and common processes could help harmonize the entire life sciences ecosystem leading to a greater understanding of the human genome and the links between genetics and human disease. In particular, this project aims at developing a Cloud-based environment for RNA-Seq and DNA-Seq data obtained in the context of breast cancer diagnosis.

RESEARCH OBJECTIVES

In order to achieve this goal, the project will accomplish the following research objectives:

1. The development of the necessary Cloud Infrastructure for data storage and processing;
2. The implementation of bioinformatics pipeline for genomics, transcriptomics and epigenomics analyses;
3. The development of a machine learning algorithm for quantitative evaluation of breast cancer based on clinically relevant gene variants and imaging techniques.

The described genomics Cloud architecture will enable a strong worldwide collaboration, enabling sharing processes and methodologies around the globe, while guaranteeing patient

data security and protection, which will make research more FAIR: findable, accessible, interoperable, and reusable[3]. The increase in data volume, complexity and data generation speed entail a progressive need of computational support. In this view, we need to rethink how biological data are generated and processed, and the genomic Cloud is able to support such movement towards a more dynamic and sophisticated way of working.

CANDIDATE EXPECTED BACKGROUND

The ideal candidate for this project is expected to have a strong background in computer science and Cloud computing, with previous experience in the field of big data analytics and processing. In addition, he/she should have an advanced knowledge of NGS-based techniques and bioinformatic pipelines, and a basic set of skills in the machine learning algorithms. Importantly, the candidate should have a good attitude for teamwork in an interactive environment, along with a very good level of written/oral scientific english.

**Supervisor(s), Lab/Group details, other additional info**

Dr. Davide Cacchiarelli (Head of Armenise/Harvard Laboratory of Integrative Genomics at TIGEM, Telethon Institute of Genetics and Medicine).

**Funding**

The proposed research activities can count on national and international funding sources, including Telethon funds, ERC starting grant, POR and AIRC institutional grants, and also research funds of NGD (Next Generation Diagnostic, s.r.l.) a spin-off company of UNINA.

**References**

1. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
2. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
3. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).