# PROPOSAL FOR FINAL MASTER PROJECT

# UPC-UB Master's Degree in Statistics and Operations Research (MESIO)

## COMPANY / INSTITUTION

Name of the company/institution: Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)
Main activity: Biomedical research
Address: Av. Granvia de L'Hospitalet, 199-203 City and postal code: 08908
Web: http://www.idibell.cat Telephone: 932607186
Contact person: Xavier Solé Acha
Charge of the contact person: Leader of the Cancer Data Science group
Contact person e-mail: x.sole@iconcologia.net

## DESCRIPTION OF THE ACTIVITY

**Name of the activity:** Design and implementation of the statistical pipeline for the SNPStats 2.0 genetic epidemiology web tool

**Description of the activity:**
SNPStats is a web-based tool for candidate-gene population genetics analyses. Created more than a decade ago, the tool has already performed over 300,000 analyses from researchers across the world, and has been cited close to 1,000 times in peer-reviewed scientific publications based on Google Scholar. Essentially, the tool requires the user to input a set of genetic variants and additional variables determined for a cohort of individuals, and performs population genetics analyses to identify whether any of the genetic variants (either independently or jointly) are associated with a binary outcome by applying population genetics methods and logistic regression modeling.

Despite its popularity, the functionality, interface, and the underlying technology of the current version of SNPStats can now be greatly improved, and the aim of our group is to develop a more modern version of the tool with added functionality, as well as improved interface and user experience. We have already designed and implemented a new web interface, and the next step is to create the new statistical engine than will communicate with the interface and perform the analyses required by the user. The project will NOT require any knowledge of web development or any other programming languages than R. The student required for this project will focus entirely on implementing the genetic epidemiology pipeline of the SNPStats 2.0 tool. More specifically, the project will involve designing and coding a statistical pipeline in R involving the following statistical tests/models, among others:

- Hardy-Weinberg tests.
- Logistic regression.
- Linear regression.
- Survival analyses.
- Interaction analyses.

The project will be developed in the context of the Cancer Data Science (CaDS) group, led by Dr. Xavier Solé. The CaDS group is part of the Cancer Prevention and Control Program of the Catalan Institute of Oncology in Barcelona, Spain.

**Period of the activity:**

Aproximate period of the activity: The project is expected to last approximately 4 months with a full time dedication.

**Wage Compensation:** Yes ☐ No ☒ Only travel expenses ☐

Total amount for the entire period of internship: 0 EUR.

**Requirements to be met by the student:**

The student will be responsible for designing and developing the pipelines and tools already available for this type of analyses. Prior knowledge of the statistical computing framework "*R*" is essential. Knowledge of biostatistics and statistical models is also required. Basic concepts of programming and algorithmic skills are highly desirable. Basic knowledge of the Linux operating system is not required but will be a plus. Knowledge of English at the reading level is essential. The student should feel comfortable working in a collaborative, multidisciplinary environment, and be motivated to deal with new challenges.

**Other observations:**

Group Leader
Xavier Solé Acha, PhD
E-mail: x.sole@iconcologia.net
Research group website: http://icoprevencio.cat/ubs/
SNPStats website: https://www.snpstats.net/snpstats/