# Machine learning for early identification of immune cell subtypes

This research internship is related to the project Plascan (Analysis and Epigenetic Recognition Of Dysbalanced Immune Cell Plasticity) that involves researchers from the department of immunity, virus and inflammation of the CRCL (PI Hector HERNANDEZ-VARGAS) and Data mining and Machine Learning team of LIRIS laboratory (PI Céline Robardet). The main objectives of the project is to propose machine learning methods to be applied on sequencing signals produced by third-generation nanopore DNA sequencers (such as MinION) to early identify immune cell subtypes.

## Context

**Immune cell plasticity defines the anti-cancer response.** Cancer cells can be either detected and destroyed by our immune system, or tolerated and left to divide and proliferate. Such immune response against cancer is coordinated by different immune cell types. Among those, T cells are known to play a key role. Specifically, the balance between T cell subtypes (e.g. Th0, Tregs, Th17) influences the final outcome of the response. Moreover, T cell subtypes are not fixed, and there is a known "plasticity" defined by transcription factors differentially active under certain conditions [1]. For example, Th17 cells have been shown to be unusually flexible in their developmental programs [2], and their deviation towards Th1 is associated with strong chronic inflammation and cancer. Similar skewing towards inflammation has been described for Treg cells.

**DNA methylation identifies cell subtypes.** A unique DNA template is able to originate the diversity of gene-to-protein programs that are particular to each cell type. Chemical modifications, such as DNA methylation provide a system to pass information through cell generations while keeping the same DNA sequence (in a so called "epigenetic" process). In mammals, DNA methylation occurs at the 5th carbon of cytosines (5mC) preceding guanines (CpG sites), and together with its recently described derivatives (i.e. 5hmC, 5fC, and 5caC), it is known to be tissue- and cell type specific [3]. While most CpG sites in the genome are stably methylated, only a small fraction of variably methylated loci seems to be informative in terms of cell identification. For example, our recent analyses show that at least 5000 CpG sites are differentially methylated between Th0 and Th17 cells (unpublished data). Similar findings have been described for every immune cell subtype, although this is probably an underestimation, due to limitations in coverage of techniques routinely used to detect 5mC.

**Third generation sequencers are game-changers.** Recently developed nanopore DNA sequencers (such as the MinION) are portable real time, long-read, low-cost devices that promise to revolutionize genomic research, motivating a faster translation of fundamental research into clinical practice. Such instruments measure the change in electric current caused by DNA transiting through a pore. It has been shown that the electrolytic current signals are sensitive to base modifications, such as 5mC. For example, using synthetically methylated DNA it was possible to train a hidden Markov model to distinguish 5mC from unmethylated C [4]. However, no algorithm is currently available to reliably identify methylation events in a non-synthetic "real life" context, such as the direct mapping of the immune cell landscape.

## Internship objectives

Given this data science context, the objective of the internship is to develop a Machine Learning approach that identifies chemical cytosine modifications from  sequencing signals obtained from immune cells. Two types of classification problems will be considered: (1) the identification of local DNA methylation by classifying cytosine nucleotides as methylated or not; (2) the identification and characterization of T cell subtypes. A neural network will be designed and used to predict the methylation based on the electrical signal, but it will also use the surrounding nucleotide sequence in order to take advantage of the structure and redundancy of the DNA to potentially correct measurement errors. This methylated cytosine nucleotide detection problem faces the challenge of classification in imbalanced data, well identified in artificial intelligence. A step further will consist to integrate prior biological knowledge into the classification model and also to analyze it in order to potentially disclose assumptions about the methylation process. In this sense, emphasis will be done in detecting methylation variations at loci already known to distinguish these cell subtypes.

## Expected profile

We are looking for an outstanding computer science student with a broad curiosity. The ideal candidate would also have the following qualities:
• A good prior knowledge in Data Mining/Machine Learning/Data Science/Big Data. You'll  do not need to already be an expert, but you need to have the basics from which to learn by yourself.
• An interest for interdisciplinary research.

If you're interested by this project, you have questions or would like to apply,
contact Stefan Duffner (stefan.duffner@liris.cnrs.fr), Marc Plantevit (marc.plantevit@liris.cnrs.fr) and Céline Robardet (celine.robardet@insa-lyon.fr), joining a CV and a few lines describing what are your skills in relation to the project.

**References**

1. Ziegler SF et al. FOXP3 and the regulation of Treg/Th17 differentiation. Microbes Infect. 2009 Apr;11(5):594-8. doi:10.1016/j.micinf.2009.04.002.

2. Basu R et al. The Th17 family: flexibility follows function. Immunol Rev. 2013 Mar;252(1):89-103. doi: 10.1111/imr.12035.

3. Ecsedi S et al. 5-Hydroxymethylcytosine (5hmC), or how to identify your favorite cell. Epigenomes 2018, 2(1), 3; doi:10.3390/epigenomes2010003.

4. Simpson JT et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017 Apr;14(4):407-410. doi:10.1038/nmeth.4184.