

**Title:** Identifying signatures of human epigenetic modifications among tissues.

**Main Supervisor:** Thomas Bataillon

**Main Supervisor email:** [tbata@birc.au.dk](mailto:tbata@birc.au.dk)

**Start date:** 01-02-2019

**End date:** 14-06-2019

### **Project description**

The number of different epigenetic landscapes for a genome may be inestimable, but we can find correlations between specific epigenetic modifications which are typically associated in concrete functions and development states. In such a way, we reduce the dimensionality of the problem making it easier to draw conclusions from the analysis of the epigenetic modifications, as well as being able to use the smaller set of correlated modifications (or "signatures") as input for predictive modelling or supervised machine learning analysis.

In order to achieve this, we need to map epigenetic modification reads (from Bisulfite-seq; DNA methylations, or ChIP-seq; histone modifications) into the human genome, specifically into genes and flanking/regulatory regions, which are the ones of interest. In this way we would obtain the counts of each epigenetic modification for different tissues. Non-negative Matrix Factorization (NMF) reveals as an ideal method for the task of finding combinatorial patterns of epigenetic modifications. We can then study the state of each epigenetic modification type in the defined loci of the tissue. From this information we would obtain the different epigenetic signatures which we will use for association and simulation analysis.

### **Learning goals:**

- The student should be able to implement methods in a pipeline capable of processing raw data into useable data.
- The student should be able to implement scripts for mapping epigenetic modifications into a genome.
- The student should be able to discuss about epigenetic modifications and their effect in gene expression.
- The student should be able to identify combinatorial associations between epigenetic marks based on ChIP-seq data.
- The student should be able to describe the identified association by the use of plots and summary statistics.
- The student should be able to use machine learning approach in order to predict future data.