**Integration analysis of ‘OMICS’ data using penalized regression methods: An application to bladder cancer**

Silvia Pineda1,2, Nuria Malats1 , Kristel Van Steen2

**Affiliations of authors**:

1 Spanish National Cancer Research Center (CNIO), Madrid, Spain.

2 University of Liege (ULg), Liege, Belgium.

There is a growing interest in combining different ‘omics’ datasets to further dissect the mechanisms of human complex disease traits. The simplest form of data integration involves two different data types (for instance, GWAS and expression data, as in eQTL analyses). The availability of more than 2 omics data types derived from the same set of individuals is rare. And when these exist, several technical and statistical hurdles need to be taken to ensure optimal power and reliable biological relevant relationships. In this work, we rely on variable selection methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) approach and the Elastic Net method. Although these are promising statistical methods presenting good properties in the context of high-throughput data, they do not provide p-values to assess statistical significance of relationships, or give a formal assessment of the overall goodness-of-fit. Therefore, we adopt a permutation-based strategy to assess significance of discovered relationships, building upon the concepts of deviance and the Mean Squared Error (MSE). Validity and utility of these methods is shown on synthetic data as well as real data from the pilot Spanish Bladder Cancer/EPICURO study (bladder cancer cases recruited in 2 hospitals in Spain in 1997-1998), while integrating gene expression, DNA methylation and genome-wide SNP data from tumor samples.