Abstract

One of the most common types of analysis in genome research is the comparison of gene expression profiles (or transcriptomics) to understand the relationship between genes (or genotype) and the phenotype. Transcriptome analysis has been traditionally conducted using microarrays and with increasing frequency since 2008 by RNA sequencing (RNA-seq). A fundamental goal in these types of studies is to identify the genes whose expression changes between different conditions, in other words, to select the most relevant variables (genes) in terms of inter-condition variability. The variable selection problem, usually known in transcriptomics as "differential expression analysis", can be addressed from the univariate or multivariate point of view, but must always take the complexity of the experimental design into account. One of the challenges that biostatisticians face when tackling this problem is the so-called "curse of dimensionality": hundreds or even thousands of variables have to be analyzed with few observations usually available. Therefore, it is essential to provide researchers with efficient statistical tools to perform this task.

A typical approach to variable selection is to test the null hypothesis of equality of average expression levels between two conditions in a gene-wise fashion and to do this repeatedly for all genes in the transcriptomics data set. However, transcriptomics may also involve multifactorial experimental designs (eg multiple treatments, several developmental states, time series...). The first part of this thesis is dedicated to the variable selection problem in multifactorial designs when multivariate methods are used to model microarray gene expression profiles. In particular, we chose the ASCA-genes multivariate technique as a starting point to propose some strategies to select differentially expressed genes in the multivariate context, that were tested under different biological scenarios.

RNA-seq technology emerged when work for this thesis was first started and now it is widely applied in transcriptomics. RNA-seq data is fundamentally different from microarray data and this has motivated the development of new statistical methods to study differential expression in this technology. Thus, the second part of the thesis is entirely focused on RNA-seq experiments. First, we develop a set of procedures to assess the quality of RNA-seq measurements, to identify the potential biases of the technology and to process the data to reduce the impact of technical noise on statistical results. Secondly, we address the variable selection problem for the two class comparison case. Given that some controversy exists on the theoretical distribution followed by RNA-seq data, we opted to investigate non-parametric data-driven techniques to overcome the limitations of parametric assumptions and propose a strategy that is efficient in controlling the false positive rate. Two methodologies, NOISeq for technical and NOISeqBIO for biological replicates, were developed and compared to the state of the art methods.