Differential expression in RNA-seq: a matter of depth

Sonia Tarazona , Fernando Garcia-Alcalde, Alberto Ferrer, and Ana Conesa
Centro de Investigacion Principe Felipe, Valencia, Spain

Next Generation Sequencing (NGS) technologies are revolutionizing genome research and in particular, their application to transcriptomics (RNA-seq) is increasingly being used for gene expression profiling as a replacement for microarrays. However, the properties of RNA-seq data have not been yet fully established and additional research is needed for understanding how these data respond to differential expression analysis. In this work we set out to gain insights into the characteristics of RNA-seq data analysis by studying an important parameter of this technology: the sequencing depth. We have analyzed how sequencing depth affects the detection of transcripts and their identification as differentially expressed, looking at aspects such as transcript biotype, length, expression level and fold-change. We have evaluated different algorithms available for the analysis of RNA-seq and proposed a novel approach -NOISeq- that differs from existing methods in that it is data-adaptive and non-parametric. Our results reveal that most existing methodologies suffer from a strong dependency on sequencing depth for their differential expression calls and that this results in a considerable number of false positives that increases as the number of reads grows. In contrast, our proposed method models the noise distribution from the actual data, can therefore better adapt to the size of the dataset and is more effective in controlling the rate of false discoveries. This work discusses the true potential of RNA-seq for studying regulation at low expression ranges, the noise within RNA-seq data and the issue of replication.