A normal mixture model to differentiate signal from noise in RNA-seq, ATAC-seq, and epigenetic mark ChIP-seq data

Mariia Karapetiants

Department of Biomedical Sciences, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

Epigenetic marks are an essential part of transcriptional control. Often located in cis-regulatory elements, they may promote or silence gene expression. Patterns of epigenetic marks may indicate, e.g., developmental stages, sequences of cell lineage differentiation, or pathological processes. Sequencing of epigenetic marks, as any other sequencing technique, is prone to technical errors, caused by amplification, fragmentation, etc, resulting in a lack of consistent signal particularly for regions with low read counts. While many studies focus on the distribution of noise and its filtering for mRNA-seq data, fewer have examined epigenetic noise, especially at specific cis-regulatory elements in the genome.

To mitigate the negative impact of noise on subsequent statistical analysis and interpretation of results, we developed a noise-filtering method suitable for epigenetic and expression data. It incorporates information from all samples simultaneously, avoiding the selection of a reference. Employing a modified Gaussian mixture model on log-transformed counts, we use an EM algorithm to optimally discriminate noise. We show the effect of filtering and subsequent normalization on the analysis of genome accessibility and histone marks data of different cell types and compare our approach with existing techniques.