Statistical Analysis of RIP-Chip Data: Identification of mRNAs Enriched by Immunoprecipitation

Lisete Sousa (lmsousa@fc.ul.pt)

Faculdade de Ciências da Universidade de Lisboa & CEAUL

Pre-mRNA splicing is an essential step in the post-transcriptional gene expression control involving protein-splicing factors like U2AF65. This protein is exported to the cytoplasm and involved in some other cellular functions. The identification of U2AF65-associated mRNAs under native conditions was performed by RIP-chip technology (RNA binding protein ImmunoPrecipitation on chip) using the Affymetrix GeneChip Human Genome U133 Plus 2.0.

When the most common methodologies for quality assessment, low level analysis (background adjustment, normalization and summarization) and identification of differentially expressed genes (DEGs), are applied to RIP-chip data, the obtained results differ. This probably happens because usually more than 20% of the mRNAs are enriched by immunoprecipitation procedures, while methods for normalization and identification of DEGs are developed believing that only a small proportion of genes (1% or 5%, say) express differently. We implement a background correction method inspired in a non-specific hybridization method used for pre-processing ChIP-chip data (Chromatin ImmunoPrecipitation on chip): linear regression models are used in each array to model the non-specific hybridization; probe intensities on the array are standardized using their predicted intensity as well as the variance of similar predicted intensities; the standardized probe intensities showed no need for further normalization, so the scores could be directly compared. It is proposed a probe set score, a probe set enrichment value and its p-value for enriched gene selection.

Data sets from the U2AF65-associated mRNAs experiment were analysed. Our method shows high accuracy when applied to Spike-In U133 dataset, which is used to benchmark methodologies for analysing Affymetrix microarrays.

Keywords: Linear models, pre-processing RIP-chip data, RNA binding proteins.