Unraveling the molecular profile of alternative transcripts through analysis of eCLIP and RNA-Seq data

Pedro Rodrigues Sousa da Cruz[1], Felipe Ciamponi[1], Katlin Massirer[1]

1 CBMEG - UNICAMP

Abstract

Genomic studies estimated that around 95% of human genes undergo alternative splicing and about 37% of them generate multiple protein isoforms, thus adding to the proteome complexity. Since pre-mRNA splicing is an essential process in mammalian cells, its failure can lead to overall or tissue-specific misregulation, possibly leading to diseases such as cancer. Although there has been a striking progress in uncovering the regulatory aspects of splicing networks, there is still relatively poor knowledge on the mechanisms that drive the occurrence of specific splicing events. RNA-biding proteins (RBPs) compose the main class of splicing regulators by binding to sets of mRNA-targets. We aim to assess the characteristics of the splicing site regions for those mRNA-targets for both control and RBPs individual knockdown cells in order to understand features that lead to specific events. To accomplish that, we built an analysis pipeline consisting of processing publically available enhanced CLIPseq (eCLIP) data from ENCODE; composing a reliable RBP-target list by filtering significant peaks; analyzing splicing patterns affected by these RBPs' knockdown from RNA-Seq data present in Genome Expression Omnibus (GEO; data also retrieved from ENCODE on the same cell lines) using FASTQC, trimmomatic, STAR, rMATS and R plotting functions; and finally building the profile of regions differentially used in splicing compared to the profile of general RBPs targets found in the first step. To perform the profiling we applied an algorithm developed by our group named BioFeatureFinder that gathers 5,498 features ranging from conservation data to physical characteristics as input and ranks them by significance. The eCLIP analysis showed the following splicing RBPs to be bound in splicing regions: TROVE2, PRPF8 (5' portion of the splicing region), SF3B4, SF3A3, U2AF1 and LARP7 (occupying the 3' portion). As for the profiling, BioFeatureFinder showed conservation to be a major aspect in RBPs target regions, additionally, these regions were found to be enriched for interactions with BUD13, SMNDC1 and EFTUD2, another splicing RBPs. Moreover, GC content and secondary structures were inferred to be important features shared by different targets under study.

Funding: CAPES, FAPESP