Computational Analysis of Intron Retention Events in Alternatively Spliced mRNA

Huffman, Raymond

Alternative splicing is an inherent gene regulatory mechanism, allowing for a single gene to code for a multitude of proteins. Intron retention (IR) is a class of alternative splicing variation that has largely eluded scientific understanding. Although it is predicted that IR is the most common form of alternative splicing, there is currently no established or computationally robust method for identifying IR events. Intron retention occurs when a region of DNA intended to be spliced out and removed from pre-mRNA is instead included in the final mRNA transcript, at times drastically modifying the final protein construct. This project aims to (1) investigate potential IR events to reveal hallmarks of the process, and (2) identify true IR events from RNA-seq data. RNA-seq samples were acquired from publicly available GEUVADIS database. Hypothesized features indicative of IR events were explored, including sequence alignment scores and the average sequencing read coverage. Additionally, relative abundance scores for Cytosine and Guanine base pairs within introns was computed. These suspect features were then assembled into a multi-dimensional matrix, dimensionality was reduced through principle component analysis (PCA) and unsupervised machine learning was leveraged to cluster samples. All computation was executed in Python, and clustering was implemented using the open source packages scikit-learn and NumPy. This analysis revealed that sequence features alone provide sufficient information to identify and differentiate IR events. This research establishes a unique framework for the identification of intron retention that does not require experimentation in the phenotype, enabling genomics researchers to elucidate why this type of splicing occurs.