Cis-regulatory G-quadruplex Motifs are Preferentially Associated with Splice Sites in the Protein-Coding Human Genome

Vanesa Getseva, Scott Frees, Paramjeet S. Bagga

School of Theoretical and Applied Science, Ramapo College of New Jersey, Mahwah, NJ, USA

April 2020

Abstract

Expression of mammalian genes involves regulated RNA splicing. Most human genes undergo alternative splicing during gene expression. As a result, the human protein-coding genome provides a rich variety of proteins with complex and diverse functions. It is estimated that up to one-fifth of human diseases are associated with altered splicing.

Our lab studies the role of cis-regulatory motifs, such as Quadruplex forming G-Rich Sequences (QGRS) in RNA processing. We focus on computationally identifying QGRS distribution patterns near splice sites in the human protein-coding genome and investigate their role in regulated splicing. Our dataset consists of 19,938 genes, 497,330 exons, 442,337 introns, and 369,995 unique splice sites based on the GRCh38 Homo sapiens assembly extracted from the Human Ensembl database. We have developed scripts in Python3 and C++, based on our previously established QGRS Mapper program, to map QGRS motifs.

Our analysis discovered a preferential association of QGRS motifs with splice sites in exons and introns. We observed differential QGRS distribution patterns between 5' and 3' splice sites. RNA QGRS motifs in the vicinity of specific splice sites may be involved in modulating splicing via interactions with regulatory proteins that bind G-rich sequences and influence splicing events. QGRS motifs were significantly more likely to overlap the alternatively spliced sites as compared to the constitutive sites, suggesting their role in regulated alternative processing. Our data shows that QGRS motifs are likely involved in influencing splicing of the human protein-coding genes on a genomic scale.

We are creating a UCSC Genome Browser Track Hub based on our mapped data to visualize the QGRS motifs and their prevalence on the human genome. Developing this freely accessible online Bioinformatics tool allows the world to be able to map QGRS motifs and analyze their distribution patterns on a genomic scale.

Keywords: Alternative Splicing, RNA Processing, Regulatory Motifs, G-quadruplexes, Post-transcriptional Regulation.