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

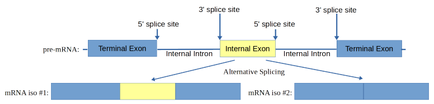
Abstract

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 19,948 genes, 451,323 exons, 406,201 introns, and 365,167 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.

Introduction

Regulated Splicing and Gene Expression

Expression of mammalian genes involves regulated RNA processing events such as splicing. Almost all human genes are thought to 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 has been estimated that up to one-fifth of the human diseases are associated with altered splicing. Therefore, studying regulation of splicing is an important area of research.



Alternative patterns of pre-mRNA splicing can lead to related yet different mRNA isoforms. Splicing patterns may vary with tissues, physiological state of cells, and development.

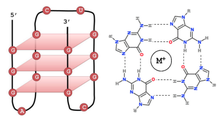
G-quadruplexes

A **G-quadruplex** is a four stranded stable 3-D structure formed by **Guanine Rich** nucleic acids. It consists of square co-planar arrays of four guanine bases known as "tetrads". G-quadruplexes play significant roles in important biological processes, human diseases, and as therapeutic targets. **Putative G-quadruplexes** can be identified as **Q**uadruplex forming **G-R**ich **S**equences (**QGRS**) using the following motif:

$G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$

x = number of guanine tetrads in the G-quadruplex

$\mathbf{y_1y_2y_3}$ = the length of the loops connecting the guanine tetrads



***A G-quadruplex structure**. **Left**: an intramolecular G-quadruplex formed by $G_3AG_3CUG_3CG_3$ RNA motif. **Right**: a G-tetrad structure.*

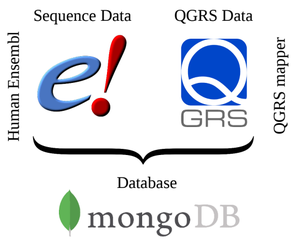
Problem and Purpose

The genome-wide role of cis-regulatory QGRS in regulated splicing is not well understood. The purpose of this investigation was to identify QGRS distribution patterns near alternative and constitutive splice sites in the protein-coding human genome, with a goal to investigate their role in regulated splicing.

Methods

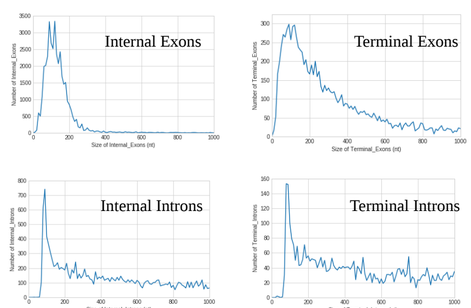
I. Homo sapiens GRCh38.85
II. Identify protein-coding transcripts (TSL≤3).
III. Map splice sites onto sequence data.
IV. Recognize alternative and constitutive splice sites.
V. Recognize exons and introns.
VI. **Relate alternative and constitutive splice sites with QGRS motifs.**

Data



An Overview of the Dataset

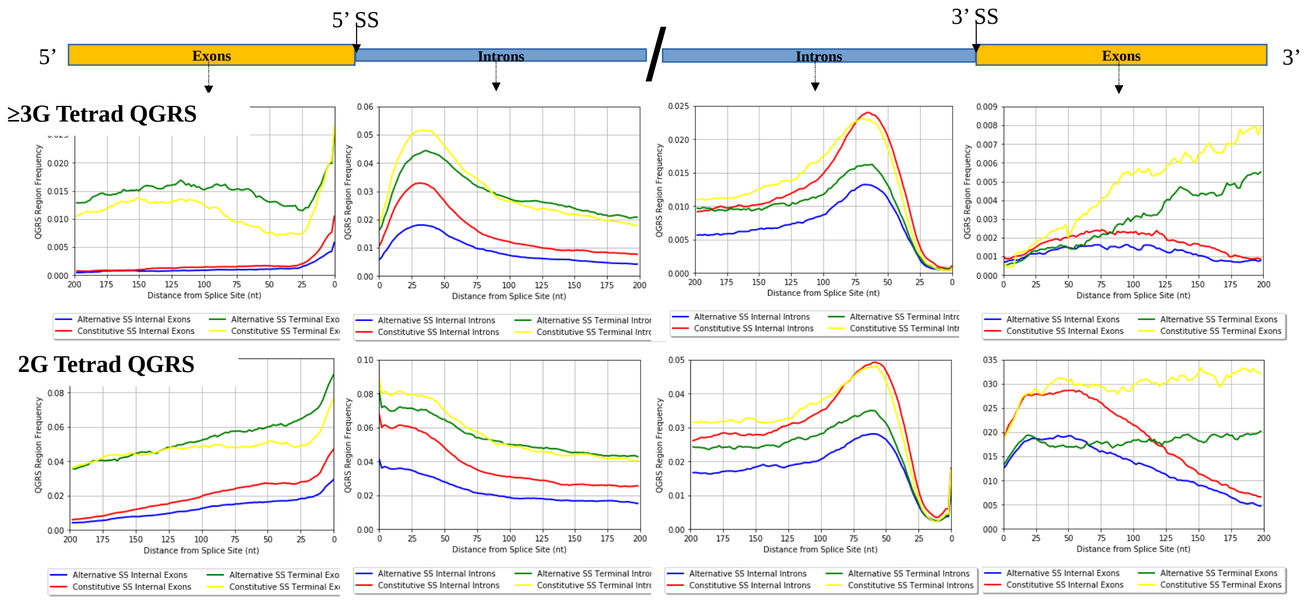
Genomic Structure	Total #
Genes	19,938
Transcripts	54,993
Exons	497,330
Internal Exons	387,344
Terminal Exons	109,986
Introns	442,337
Internal Introns	336,497
Terminal Introns	105,840
Unique Splice Sites	369,995
Alternative 5' Splice Sites	103,427
Constitutive 5' Splice Sites	83,473
Alternative 3' Splice Sites	98,556
Constitutive 3' Splice Sites	84,539



A majority of internal exons are 75 to 200 nucleotide bases long while terminal exons tend to be much larger. Internal as well as terminal introns are generally large.

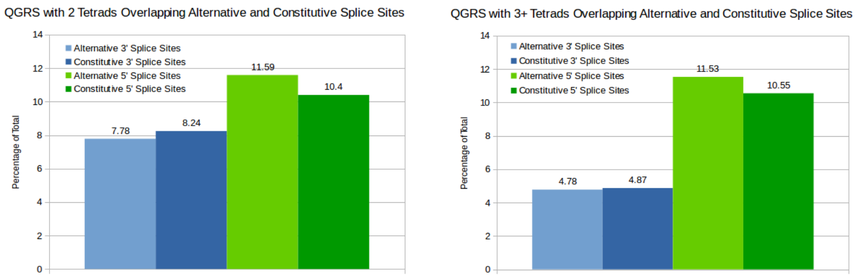
Results

Distribution of QGRS Motif Regions near Splice Sites in the Human Protein-Coding Genome



***QGRS motifs are more prevalent in the vicinity of 5' and 3' splice sites, strongly suggesting their roles in splicing.** The 2G-tetrad and ≥3G-tetrad QGRS show similar distribution patterns. **QGRS region distribution patterns around the splice sites are similar in the terminal as well as internal introns and exons. However, terminal introns and exons exhibit overall high frequencies of the QGRS occurrence. There is a high concentration of QGRS in the introns between 20-50 nucleotides from the 5' splice site, suggesting their role in the initial stages of the splicing process**. A majority of the QGRS regions were found within 50 to 100 nucleotide bases from the 3' splice site of the internal introns. **The preferential high density association of intronic QGRS within <100 nucleotides from the 3' splice site places them in the area of U2 snRNP and regulatory proteins binding region, suggesting a potential mechanism of their involvement in splicing.** (A large number of exons of length 100 nt or higher and introns of 200 nt or larger length were used for this analysis. The data represents analysis of 200 nucleotides on either side of 5' and 3' splice sites in exons and introns. The data represents protein-coding transcripts with TSL≤2).*

QGRS Motifs Overlapping Splice Sites



Alternative 5' splice sites are more likely to be overlapped by QGRS as compared to 5' constitutive sites. The same relationship was not detected for 3' splice sites.** (A total of 168,012 constitutive splice sites and 201,983 alternatively spliced sites were analyzed). We believe that when a G-quadruplex overlapping the splice site is formed the splice site becomes unavailable. The G-quadruplex needs to be resolved for alternative splicing to occur. Many cellular proteins are known to help form or resolve the G-quadruplex structure. **Our data suggests that highly stable G-quadruplex motifs can help regulate alternative splicing at the 5' splice sites.

Selected Splice Sites Associated with QGRS Motifs

Gene	Biological Relevance	Splice Site (nt position)	QGRS Distance from SS	QGRS Motif	QGRS-SS Relationship
NIPA1 (ENSG00000170113)	Mg transporter. Nervous system development.	5' (22,786,248)	Overlap	GGTGAGTTGGGCCGGCCTCCGCTGG	QGRS overlapping Alt 5'-SS
MIB2 (ENSG00000197530)	Ubiquitination of proteins in the Notch signaling pathway.	3' (1,623,774)	-63 to -36 nt	GGGCAGGCGGGACGGGCAGGACCCGGG	QGRS near 3' Alt SS
PLEKHN1 (ENSG00000187583)	Pleckstrin homology domain containing N1. Apoptosis.	3' (970,879)	-70 to -43 nt	GGAGAGGGGCCTGCCCTGTGGCTGCGG	QGRS near 3' Alt SS

Conclusions

- We found that QGRS motifs are more prevalent in the vicinity of 5' and 3' splice sites, strongly suggesting their roles in splicing.
- QGRS region distribution patterns around the splice sites are similar in the terminal as well as internal introns and exons. However, terminal introns and exons exhibit overall high frequencies of the QGRS occurrence.
- There is a high concentration of QGRS in the introns between 20-50 nucleotides from the 5' splice site, suggesting their role in the initial stages of the splicing process.
- A preferentially high density association of intronic QGRS within <100 nucleotides from the 3' splice site places them in the area of U2 snRNP and regulatory proteins binding region, suggesting a potential mechanism of their involvement in splicing.
- We found that 5' alternative splice sites are more likely to be overlapped by QGRS as compared to 5' constitutive sites.
- **Our data suggest that QGRS motifs are likely to be involved in influencing splicing of the human protein coding genes on a large scale. We propose a role of G-quadruplex RNA binding proteins in regulated splicing.**